Probabilistic Counterfactuals: Semantics, Computation, and Applications

Alexander A. Balke
Judea Pearl

Final Technical Report
Submitted to
U.S. Air Force / Office of Scientific Research
F-49620-93-1-0421
7/1/93 – 6/30/96

This report is based on A. Balke's dissertation submitted to UCLA in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Computer Science.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	2/5/97	FINAL TECHNICAL REPORT (7/1/93 - 6/30/~~96~~ 95)

4. TITLE AND SUBTITLE
TITLE: Dynamic Networks Techniques for Autonomous Planning and Control
SUBTITLE: Probabilistic Counterfactuals

5. FUNDING NUMBERS
G - F49620-93-1-0421

6. AUTHOR(S)
Professor Judea Pearl

AFOSR-TR-97 0629

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UCLA
Computer Science Department
4532 Boelter Hall
Los Angeles, CA 90095-1596

8. PERFORMING ORGANIZATION REPORT NUMBER
A87-1670F-01
442510-22525

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Air Force / Office of Scientific Research
110 Duncan Avenue, Suite B115
Bolling Air Force Base
Washington, DC 20332-0001

nm

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release. Distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

We have reformulated Bayesian networks as carriers of causal information. The result is a more natural understanding of what the networks stand for, what judgments are required in constructing the network and, most importantly, how actions and plans are to be handled within the framework of standard probability theory. Starting with functional description of physical mechanisms, we were able to derive the standard probabilistic properties of Bayesian networks and to show:

* how the effects of unanticipated actions can be predicted from the network topology,
* how qualitative causal jedgments can be integrated with statistical data,
* how actions interact with observations, and
* how counterfactuals sentences can be formulated and evaluated.

14. SUBJECT TERMS
Keywords: Causation, counterfactuals, Bayesian networks

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
unclassified	unclassified	unclassified	SAR

ABSTRACT

Counterfactual conditionals of the form "If A were true, then C" are commonly used to express generic, law-like relationships. This dissertation provides formal semantics for interpreting such conditionals, as well as computational methods for answering queries of the form "Find the probability of C if A were true, given that A is in fact false." Here, generic knowledge is represented as a network of causal relationships among variables of interest, while specific occurrences are represented as instantiations of those variables. The counterfactual antecedent A is interpreted as a local, hypothetical change induced by forces external to the system. Counterfactual probabilities are computed using standard evidence propagation in two loosely coupled Bayesian networks — one corresponding to the factual world, the other to the counterfactual — where the probabilities are defined over the causal mechanisms governing the domain. When such probabilities are not available, we develop methods for computing either bounds on the counterfactual probabilities or qualitative beliefs, i.e., order-of-magnitude abstractions of standard probabilities.

We then demonstrate the usefulness of our formulation in application areas where counterfactual reasoning is essential but considered difficult, if not impossible, to compute. First, we examine experimental studies in which subjects do not comply perfectly with treatment assignment, thus violating the tenets of randomized experimentation. We show that it is possible in such studies to derive informative bounds on treatment efficacy, tighter than any yet reported in the statistical or the epidemiological literature. Next, we address the problem of determining legal responsibility (e.g., whether the defendant is liable for the plaintiff's injuries). Although counterfactual assertions in this domain cannot be evaluated using conventional statistical analysis, under our formalism they can be assigned meaningful probability intervals. In the areas of econometrics and the social sciences, the formalism allows coherent evaluation of policies involving the control of variables that, prior to enacting a given policy, were influenced by other variables in the system. Finally, in the area of artificial intelligence, the formulation provides a computational model for interpreting counterfactual utterances, answering counterfactual queries, and evaluating actions and plans.

TABLE OF CONTENTS

List of Figures

List of Tables

Part I

Semantics

CHAPTER 1

Counterfactuals

1.1 Introduction

A counterfactual conditional has the form

If A were true, then C would have been true

where A, the counterfactual antecedent, specifies an event that is contrary to one's real-world observations, and C, the counterfactual consequent, specifies a result that is expected to hold in the alternative world where the antecedent is true. A typical instance is "If Oswald were not to have shot Kennedy, then Kennedy would still be alive" which presumes the factual knowledge of Oswald's shooting Kennedy, contrary to the antecedent of the sentence.

The majority of the philosophers who have examined the semantics of counterfactual sentences [Goo83, HSP81, Nut80, cou93] have resorted to some form of logic based on worlds that are "closest" to the real world yet consistent with the counterfactual's antecedent. Ginsberg [Gin86], following a similar strategy, suggested that the logic of counterfactuals could be applied to problems in planning and diagnosis in Artificial Intelligence. The few other papers in AI that have focussed on counterfactual sentences (e.g., [Jac89, PAA91, Bou92, Gra91]) have mostly adhered to logics based on the "closest world" approach.

In the real world, we seldom have adequate information for verifying the truth of an indicative sentence, much less the truth of a counterfactual sentence. Except for the small set of relationships between variables which can be modeled by physical laws, most of the relationships in one's knowledge base are nondeterministic. Therefore, it is more practical to ask not for the truth or falsity of a counterfactual, but for one's degree of belief in the counterfactual consequent given the antecedent. To account for such uncertainties, [Lew76] has generalized the notion of "closest world" using the device of "imaging"; namely, the closest worlds are assigned probability scores, and these scores are combined to compute the probability of the consequent.

Missing from the "closest world" approach is a precise specification of the closeness measure itself, which is critical to the analysis of counterfactuals. More specifically, it does not tell us how to encode distances in a way that would (1) conform to our perception of causal influences and (2) lend itself to economical machine representation. This dissertation will provide a concrete explication of the closest world approach, one that satisfies the two requirements above.

The target of this investigation are counterfactual queries of the form:

> If A were true, then what is the probability that C would have been true, given that we know B?

The proposition B stands for the actual observations made in the real world (e.g., that Oswald did shoot Kennedy and that Kennedy is dead) which are made explicit to facilitate the analysis.

Counterfactuals are intertwined with notions of causality: We do not typically express counterfactual conditionals without assuming a causal relationship between the counterfactual antecedent and the counterfactual consequent. For example, we can safely state "If the sprinkler were on, the grass would be wet", but the contrapositive form of the same sentence in counterfactual form, "If the grass were dry, then the sprinkler would not be on", strikes us as strange, because we do not think the state of the grass has causal influence on the state of the sprinkler. Likewise, we do not state "All blocks on this table are green, hence, had this white block been on the table, it would have been green". In fact, we could say that people's use of counterfactual conditionals is aimed precisely at conveying generic causal information, uncontaminated by specific, transitory observations, about the real world. Observed facts often do reflect strange combinations of rare eventualities (e.g., all blocks being green) that have nothing to do with general traits of influence and behavior. The counterfactual sentence, however, emphasizes the law-like, necessary component of the relation considered. It is for this reason, we speculate, that we find such frequent use of counterfactuals in ordinary discourse.

The importance of equipping machines with the capability to answer counterfactual queries lies precisely in this causal reading. By making a counterfactual query, the user intends to extract the generic, necessary connection between the antecedent and consequent, regardless of the contingent factual information available at that moment.

Although some philosophers consider the analysis of counterfactuals where no causal information is available (e.g., the "All blocks on the table are green"

example), these will not be treated in this dissertation. The interpretation of counterfactuals presented here relies on a strict separation of generic background causal knowledge and transient observations of the world. The transient observations (e.g., "All blocks on the table are green") may not be used as inference rules; only the generic causal knowledge may be used for inferring beliefs from observations. [Goo83] stresses the importance of distinguishing causal information from observed facts:

> Though the supposed connecting principle is indeed general, true, and perhaps even fully confirmed by observation of all cases, it is incapable of sustaining a counterfactual because it remains a description of accidental fact, not a law. The truth of a counterfactual conditional thus seems to depend on whether the general sentence required for the inference is a law or not. If so, our problem is to distinguish accurately between causal laws and casual facts.

Because of the tight connection between counterfactuals and causal influences, any algorithm for computing counterfactual queries must rely heavily on causal knowledge of the domain. This leads naturally to the use of probabilistic causal networks, since these networks combine causal and probabilistic knowledge and permit reasoning from causes to effects as well as, conversely, from effects to causes. This representation also reflects the separation of causal knowledge from transient observations: causal knowledge is represented by the structure of the network and its parameterization, while observations are represented by the instantiation of nodes within the network.

To emphasize the causal character of counterfactuals, we will adopt the interpretation in [Pea93c], according to which a counterfactual sentence "If A were true, then B would have been true" states that B would prevail if A were forced to be true by some unspecified intervention that is exogenous to the other relationships considered in the analysis. This *intervention-based interpretation* does not permit inferences from the counterfactual antecedent towards events that lie in its past. For example, the intervention-based interpretation would ratify the counterfactual

> If Kennedy were alive today, then the country would have been in a better shape

but not the counterfactual

> If Kennedy were alive today, then Oswald would have been alive as well.

The former is admitted because the causal influence of Kennedy on the country is presumed to remain valid even if Kennedy became alive by an act of God. The second sentence is disallowed because Kennedy being alive is not perceived as having causal influence on Oswald being alive. The information intended in the second sentence is better expressed in an indicative mood:

> If Kennedy was alive today then he could not have been killed in Dallas, hence, Jack Ruby would not have had a reason to kill Oswald and Oswald would have been alive today.

This interpretation of counterfactual antecedents, which is similar to Lewis' [Lew79] *Miraculous Analysis*, contrasts with interpretations that require that the counterfactual antecedent be consistent with the world in which the analysis occurs. The set of closest worlds delineated by the intervention-based interpretation contains all those which coincide with the factual world except on possible consequences of the intervention. The probabilities assigned to these worlds will be determined by the relative likelihood of those consequences as encoded by the causal network.

Finally, the counterfactuals that may be analyzed within the context of this dissertation are limited in terms of the form of the antecedent and the types of causal relationships. Counterfactual antecedents will be limited to conjunctive clauses. For example we will not consider the veracity of the following counterfactual conditionals:

> If Bizet and Verdi had been compatriots, Bizet would have been Italian.
> If Bizet and Verdi had been compatriots, Verdi would have been French.

because, Bizet and Verdi being compatriots would be defined as, Bizet and Verdi are Italians, or Bizet and Verdi are French, or Bizet and Verdi are ..., which is a disjunction of conjunctive clauses.

In addition, we will not consider counterfactual conditionals that are counterlegals; a counterlegal is defined as a counterfactual conditional where the antecedent is impossible (e.g., violates some strict law). For example, "if this circle

were also a square, ...". Our analysis always assumes that the counterfactual antecedent is conceptually compatible (although its normal coincidence in the world may be infinitesimally rare) with the history prior to the antecedent. In other words, it must always be possible that exceptions exist to every rule in the model impinging on the counterfactual antecedent variables.

1.2 Firing Squad Example

To illustrate the intervention-based interpretation of counterfactuals, consider a firing squad with several riflemen (one called Bob) and a Captain who gives a signal to either shoot or release a prisoner charged with treason. The behavior of these agents is as follows:

- The Captain waits for the court decision.
- Bob typically fires his rifle if and only if the Captain gives the signal to shoot.
- The Traitor typically dies if and only if the Captain gives the signal to shoot or Bob fires his rifle.

Note that if the Captain gives the signal to shoot and Bob does not fire, the traitor will typically die as a result of the other riflemen shooting, but these intermediate causes will not be made explicit in this story in order to keep the model simple.

The generic causal structure that reflects this description may be represented by the structure in Figure 1.1. The three variables C, B, and T have the following domains:

$$c \in \left\{ \begin{array}{l} c_0 \equiv \text{Captain gives the signal to release the traitor.} \\ c_1 \equiv \text{Captain gives the signal to shoot the traitor.} \end{array} \right\}$$

$$b \in \left\{ \begin{array}{l} b_0 \equiv \text{Bob does not fire his rifle.} \\ b_1 \equiv \text{Bob fires his rifle.} \end{array} \right\}$$

$$t \in \left\{ \begin{array}{l} t_0 \equiv \text{Traitor dies.} \\ t_1 \equiv \text{Traitor lives.} \end{array} \right\}$$

Now consider the following discussion between two prison guards (Scott and Dave) who looking from a window at the jail could only see that Bob fired his

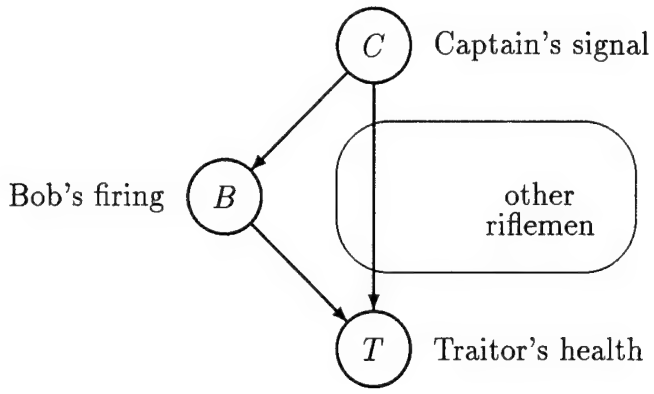

Figure 1.1: *Causal structure reflecting the influence that the Captain's signal has on Bob's firing and the Traitor's health, and the direct influence that Bob's firing has on the Traitor's health.*

rifle ($b = b_1$):

Dave:	The Captain must have given the signal to shoot, or Bob would not have fired his rifle.
Scott:	That Traitor's body must be riddled with bullets!
Dave:	Yep. If Bob were not to have fired, the Traitor would still have died.
Scott:	Ha! If Bob were not to have fired, the Captain must not have given the signal to fire, and none of the other riflemen would have fired. Therefore, the Traitor would still be alive.
Dave:	No. If Bob were not to have fired despite the Captain's signal, the other riflemen would still have fired, and the Traitor would be dead.

In the fourth sentence, Scott tries to explain away Dave's conclusion by claiming that Bob's not firing would be evidence that the Captain gave the signal to release the Traitor which would imply that none of the riflemen fired. Scott, however, analyzed Dave's counterfactual conditional in the indicative mood by imagining that he had observed Bob not firing his rifle; this allows him to use the observation for abductive reasoning. But Dave's subjunctive counterfactual

conditional should be interpreted as leaving everything in the past as it was (including conclusions obtained from abductive reasoning from real observations) while forcing variables to their counterfactual values. This is the gist of his last statement.

This example demonstrates the plausibility of interpreting the counterfactual statement in terms of an external intervention causing Bob to not fire, regardless of all other prior circumstances. The only variables that we would expect to be impacted by the counterfactual assumption would be the descendants of the counterfactual variable; in other words, the counterfactual value of Bob's firing does not change the belief in the Captain's signal from the belief prompted by the real-world observation.

The claim that the intervention-based interpretation of counterfactual antecedents should be adopted for the analysis of counterfactual conditionals in general is a controversial position. This interpretation does not cover all linguistic usages of counterfactuals; however, it does provide clear semantics and a precise computational formalism for analyzing counterfactuals given a causal description of the world. The results of this analysis provide useful information about the effects a localized change to a single variable would have on the world. In contrast, it is not clear how useful other nonintervention-based interpretations of counterfactuals are, because they imply nothing about control of ones environment. For example, some counterfactual interpretations will conclude that if Bob were not to have fired, then the Traitor would still be alive; however the Traitor's wife is not going to bribe Bob not to fire, because she knows that such an intervention will not prevent her husband from being executed. It is not the goal of this dissertation to provide a model of all linguistic usages of counterfactuals, but to provide an interpretation that lends itself to meaningful application.

1.3 Previous work

Counterfactual conditionals have been extensively studied by the philosophy community over the last twenty-five years. Of the more compelling research has been the work of Stalnaker and Lewis, from which possible-world semantics have been developed. These semantics formally describe how a closeness (or similarity) measure between worlds can be used to evaluate one's belief in a counterfactual conditional. Most research has focussed on logical inferences, but some has concerned itself with the probabilistic evaluation of counterfactuals. The 1986 paper by Matt Ginsberg injected this important topic into the Artificial Intelligence

community, where for the most part, the concentration has been on logical rather than probabilistic formalisms. What has been most lacking in this work is a precise specification of similarity between worlds; and this is paramount for practical application of the possible-world semantics.

One popular proposal has been that counterfactual conditionals may be analyzed by applying belief revision techniques, where possibly contradictory information is added to the knowledge base and information is retracted in order to bring about a consistent set of knowledge [Dal88, GM94, Gin86]. However, the intervention-based interpretation of counterfactual antecedents proposed in this dissertation is clearly not consistent with belief revision, because one may not reason abductively from the new information added to the knowledge base.

The remainder of this section will review some important contributions in the study of counterfactual conditionals.

1.3.1 Lewis' closest-world semantics

Lewis' closest-world semantics [Lew76] provides an intuitive interpretation to the analysis of counterfactual conditionals. Just find the world most similar to our observed world, such that the counterfactual antecedent holds true; if the counterfactual consequent holds true in that most similar world, then the counterfactual conditional is said to hold true.

In more detail, all worlds are first ordered relative to the observed world, which results in progressively distant "spheres of similarity" surrounding the observed world. This is graphically represented in Figure 1.2. Of interest is the closest sphere in which there exist worlds where the counterfactual antecedent A holds true. Within this set, the relative preponderance of worlds where the counterfactual consequent C holds true is evaluated, leading to a belief in the counterfactual consequent given the counterfactual antecedent.

In [Lew76], Lewis proposed a method for evaluating the probability of Stalnaker's conditionals ($A > C$) by *imaging* the set of possible worlds with respect to the antecedent A. Assume $P(w)$ is the distribution over all worlds conditioned on our partial observation of the world. For each world w, there is an imagined world w_A that is most similar to w among those worlds where the counterfactual antecedent A holds true. The probability of the worlds imaged on A (P_A) is then evaluated as

$$P_A(w') = \sum_{w:w_A=w'} P(w)$$

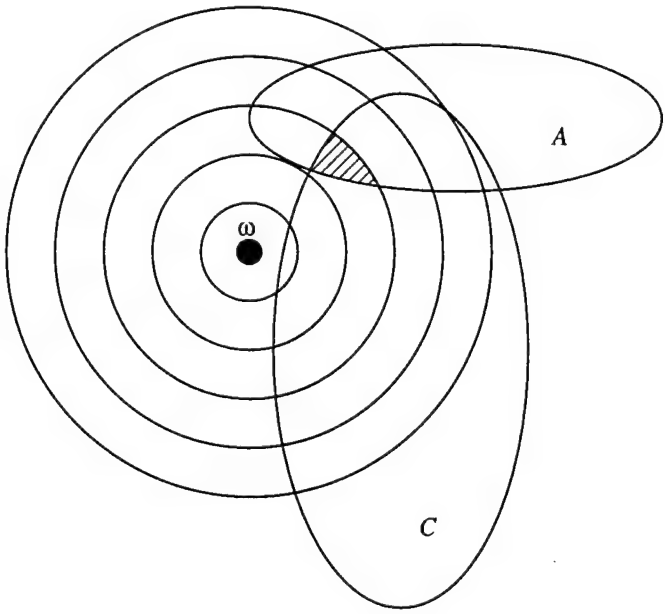

Figure 1.2: *Graphical representation of Lewis' closest-world semantics. Each circular region corresponds to a set of worlds where each world is equally similar to w. These regions are called* spheres of similarity. *The hashed region represents the set of closest worlds where the counterfactual antecedent A holds true and the counterfactual consequent holds true.*

Lewis refers to this new distribution over all worlds as the "image of P on A."

The probability of Stalnaker's conditionals $P(A > C)$ is then evaluated

$$\begin{aligned} P(A > C) &= P_A(C) \\ &= \sum_{w:w \models C} P_A(w) \end{aligned}$$

Missing from this discussion of Stalnaker conditionals, though, is a precise formulation of closest worlds (which is crucial to imaging worlds and their associated probability distribution). In Chapter 2 we present a formalism for evaluating counterfactual conditionals that is consistent with Lewis' formalism for evaluating Stalnaker conditionals via imaging. In order to make the formalism concrete, though, the notion of closest worlds must be formalized, and this will be accomplished by turning to the causal structure of the world and interpreting the counterfactual antecedent as an external intervention that forces the antecedent to be true.

1.3.2 Ginsberg

[Gin86] introduced the study of counterfactuals to the Artificial Intelligence community as an important facet of commonsense reasoning, and discussed several application areas that could benefit from the analysis of counterfactual conditionals.

Ginsberg presented a syntactic interpretation of Stalnaker's closest-world semantics. In his formulation, a world is specified by a set of logical statements S, and a closest world to S where a is counterfactually true is given by a maximal subset S' of S such that S' does not imply a. A counterfactual conditional $a > c$ is then accepted if c is true in all maximal subsets S'. In order to incorporate domain-dependent information to this strictly semantic interpretation, Ginsberg suggests the use of a "badworld" predicate to explicitly eliminate worlds from consideration. In addition, a partial order may be specified over all subsets of S to extend the set inclusion measure of closeness. In generating possible worlds, Ginsberg suggests (in the context of combinatorics) that rules of implication should not be reversible; however, it is not clear whether this is based on the belief that an implication represents a causal relationship which is not transient.

In general, though, a syntactic intepretation of closeness of worlds based on set inclusion does not reflect our understanding of causal relationships in the world. Suppose that somebody lined up 26 dominoes on end, and then tipped

the first domino towards the second domino, creating a chain reaction that finally toppled all the dominoes. Let the standing state of these dominoes be represented by the variables A, B, C, ..., Z, where $a_0, b_0, \ldots, z_0$ indicate that dominoes fell, while $a_1, b_1, \ldots, z_1$ indicate that dominoes stood. In a syntactic interpretation of counterfactuals such as that proposed by Ginsberg, one's might model their knowledge of the world by the propositions:

$$
\begin{array}{c}
a_0 \longrightarrow b_0 \\
a_1 \longrightarrow b_1 \\
b_0 \longrightarrow c_0 \\
b_1 \longrightarrow c_1 \\
\vdots \\
y_0 \longrightarrow z_0 \\
y_1 \longrightarrow z_1 \\
a_1 \\
b_1 \\
\vdots \\
z_1
\end{array}
$$

Consider the counterfactual query, "If domino B were not to have fallen, would domino Z still have fallen?" Intuitively, we reason that if B had not fallen, then there would have been no impetus to continue the chain reaction from domino C to Z. Therefore, Z would not have fallen. Under our intervention-based interpretation we would still state that A would have fallen, because we are considering the world where B was forced to stand by some intervention, e.g., domino B was nailed down. This world is given by $\{a_0, b_1, c_1, d_1, \ldots, z_1\}$.

The syntactic approach, though, does not make use of this causal information necessary for reaching the intuitive conclusion. This approach pursues the world that retracts the fewest propositions in the above set. The two closest worlds (we only show the fallen state of each domino) are given by $\{a_0, b_1, c_0, d_0, e_0, \ldots, z_0\}$ (b_0, $a_0 \rightarrow b_0$, and $b_1 \rightarrow c_1$ are retracted) and $\{a_1, b_1, c_0, d_0, e_0, \ldots, z_0\}$ (a_0, b_0, and $b_1 \rightarrow c_1$ are retracted). Clearly there is a disconnect between these results and intuition coming from our causal knowledge, because the syntactic approach does not distinguish transient observations from generic causal relationships.

If we do follow the suggestion that statements of implication should not be retracted, then there is only one closest world, $\{a_1, b_1, c_1, d_1, e_1, \ldots, z_1\}$, which leads

us to the intuitive belief that the last domino would not have fallen. However, there is a distinction between the results produced by this approach and that produced by our intervention-based interpretation: whether the counterfactual antecedent abductively implies some change to its causal effects (assuming that implication is linked to causality in the syntactic interpretation).

Ginsberg claims that counterfactuals should not be tied too closely to the notion of causality. Referring to the counterfactual conditional, "If John had koplic spots, he'd have measles." he states "... it is difficult to imagine how counterfactual implication can capture a causal relation that remains asymmetric even in this case." Under our intervention-based interpretation, this counterfactual would not hold true; it is possible that another mechanism may be found for generating koplic spots, and koplic spots do not cause measles on their own. Of course, if we observe koplic spots, we will infer that the subject has measles, but this is not the nature of a causal counterfactual conditional.

1.3.3 Simon and Rescher

Simon and Rescher discussed the analysis of causal counterfactual conditionals in [SR66]. This work is important for its formulation of counterfactuals within a causal system, and its distinction it makes between generic causal knowledge and transient observations.

They propose that when the counterfactual antecedent is included into the knowledge base, inconsistencies in one's knowledge must be retracted without violating any causal relationships ("We might of course give up the law (L), but this course is obviously undesirable."). In order to determine which knowledge is retracted, each variable is assigned to a modal category according to its distance down the causal chain from the exogenous variables. The higher the modal category, the more succeptible the variable's value is to retraction.

While Simon and Rescher choose to uphold all laws in one's model of the world, our intervention-based interpretation severs the causal link between the antecedent variables and their modelled set of causal influences. Consider Simon and Rescher's wheat growing example, where fertilizer (F) and rain (R) influence the wheat crop (W), while the wheat crop and population (N) influence the wheat price (P). Figure 1.3 shows this causal structure.

Given the above causal structure, Simon and Rescher's counterfactual conditional, "If the wheat crop had been smaller last year, the price would have been higher" is consistent with our intervention-based interpretation; however, they

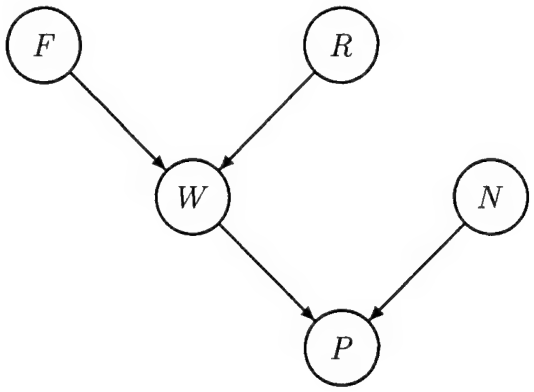

Figure 1.3: *Simon and Rescher's structure representing the causal relationships between fertilizer (F), rain (R), wheat crop (W), population (N), and wheat price (P).*

find the following statement "perfectly idiomatic": "If the wheat crop had been smaller last year, there would have been either less rain or less fertilizer applied." This contrasts with our interpretation which leaves our belief in the rain and fertilizer amounts unchanged. Simon and Rescher's interpretation fits an analysis where the antecedent is considered to be a passive observation, e.g., in a similar world where we would have observed a smaller wheat crop, either there was less rain or less fertilizer was applied. However, this analysis does not necessarily tell us the causal influence that a change in wheat crop would have had on the world, if there was another path from rain or fertilizer to the wheat price, because the value for variables preceding the counterfactual antecedent may still affect variables that are descendants of the antecedent variable.

1.4 Applications

In this section the importance of causal counterfactual reasoning will be emphasized by describing some of the tasks that benefit from such analysis. The common occurrence of counterfactual statements in everyday human discourse is a clear tipoff that counterfactuals are an integral part of human communication. Asides from the efficiency gained in communication, formal evaluation of counterfactuals is important to system design, fault diagnosis, liability litigation, policy analysis, etc. Some of these are mentioned in [Gin86].

1.4.1 Communication

Counterfactuals are a prevalent aspect of daily communication between humans, which is interesting, because very often the conditional statement is made after an irreversible event has occurred. For example, suppose that little Johnny pulled on Sarah's pigtail followed by Sarah dumping her milk shake on Johnny's head. Johnny, totally surprised, turns to his mother and cries out innocently and indignantly that Sarah has done something terrible. Johnny's mother, having observed the whole scene, calmly explains to Johnny, "If you had not pulled on Sarah's pigtail, then she would not have dumped her milk shake on you."

This counterfactual conditional is useless to Johnny at this point in time; it will not bring about a plan to get clean, and it will not allow Johnny to exact retribution. So what is his mother's point in making this statement? Precisely for conveying information to Johnny about the causal relationship between pulling Sarah's hair and Sarah's subsequent actions. It is the mother's belief that this information will allow Johnny to formulate a belief system that will hopefully discourage him from pulling Sarah's hair (at least when he does not want to suffer the consequences).

This information tells Johnny that when everything else is held fixed, that a local change to Johnny's hair pulling would evoke a change to the outcome. This is more informative to Johnny then the statements, "If you pull Sarah's hair, then she punishes you" and "If you do not pull Sarah's hair, then she will not punish you." This might not convey the same information to Johnny; he may interpret this to mean that the same situation where he would pull Sarah's hair is the same situation where Sarah is going to pour her milk shake on him, in which case he might as well go ahead and yank her hair. Thus, we see that the counterfactual conditional conveys the isolated causal effect of Johnny's hair-pulling on Sarah's response, informing Johnny that his decision whether to pull Sarah's hair will have influence on Sarah's reaction.

1.4.2 Liability litigation

Frequently, the analysis of counterfactual conditionals is required in the determination of liability in legal cases. A plaintiff might claim that a defendant's action or product has inflicted damages on their person or property, and the court must analyze the following types of questions. "If the plaintiff had not been exposed to the product, would the plaintiff still have developed his current illnes?" Or, "if

the defendant had not conspired to fix prices with the other manufacturers of widgets, would the plaintiff not have lost his business?" To answer these questions, the court ponders how a local change to the circumstances of the plaintiff (e.g., preventing the defendant's action, or removing the product from the plaintiff's environment) would have effected his welfare differently than actually occurred. If the court decides that the local change would have prevented the plaintiff from suffering financial or personal injury, then the court would find the defendant liable for those damages.

In Chapter 7 we will discuss cases where the analysis of counterfactuals involves statistical models, and we will present a hypothetical case using the partial-compliance model of Chapter 5 to demonstrate that a court must apply counterfactual probabilities in order to guarantee proper determination of liability in product-safety litigation.

1.4.3 Policy analysis

In the clinical study of new drug treatments, researchers wish to determine whether or not a particular drug will improve the overall rate of recovery of subjects within the patient population. Subject's from the population are randomized into one of two treatment groups and their treatment response is measured at the end of the study. However, the study is seldom perfect: patient's remove themselves or are removed from the study; patient's do not comply with their treatment assignment; and exogenous influences confound the results. Given the data from the study, the researchers wish to answer the following counterfactual query: "If the patient population were uniformly treated with the drug under study, would the overall recovery rate of subjects in the population have been higher than if the population were uniformly given a placebo?"

This application of counterfactual probabilities will be explored in depth in Chapters 5 and 6. In addition an example demonstrating economic policy analysis using linear structural equation models will be presented in Chapter 8.

1.5 Contributions

The principle contributions in this dissertation consist of:

- Specification of knowledge representation necessary for adequately analyzing counterfactual conditionals.

- Proposal of unambiguous semantics for interpreting the meaning of counterfactual antecedents in terms of intervention by an external force/action.
- Technique for evaluating counterfactual probabilities when a functional model of a domain is provided.
- Method for evaluating bounds on counterfactual probabilities when a probabilistic distribution is only available for observable variables, i.e., a functional model is not known. A program is available for deriving closed-form bounds when the counterfactual probability may be expressed as a linear combination of terms from the response-function distributions.
- Formulas for evaluating counterfactual distributions when the domain is modelled by linear structural equations.
- Derivation of strict bounds on average treatment effects from experimental studies involving partial compliance.

1.6 Overview

Part II of this dissertation is concerned with the theoretical and computational aspects related to the evaluation of counterfactual probabilities. In this first chapter the study of counterfactuals has been motivated and the intervention-based interpretation of counterfactuals — adopted for this research — has been introduced. In Chapter 2, a formal representation of knowledge facilitating the analysis of counterfactual probabilities will be described, and an algorithm for computing these probabilities will be developed. Counterfactual probabilities may only be uniquely identified when the background knowledge is described by a functional model. In addition, formulas are derived for evaluating counterfactual distributions when background knowledge is given by structural equation models with normally distributed disturbances. In Chapter 3, we demonstrate how bounds on counterfactual probabilities may be computed/derived, when the general knowledge of the world is described by a causal structure and conditional probabilities over the observable variables. Chapter 4 discusses the evaluation of counterfactual conditionals when beliefs are represented by order-of-magnitude abstractions of probabilities.

In Part III, the evaluation of counterfactual probabilities will be demonstrated in a set of applications. The most appealing of these applications is the evaluation of treatment effects in studies where subjects are randomly assigned treat-

ment, but do not necessarily comply with this assignment. This task has been studied by [EF91] with a concentration on continuous values of treatment consumed, and [Man90] has derived nonparametric bounds on treatment effects for generalized treatment domains. In Chapter 5, we derive the tightest-possible assumption-free bounds on treatment effects from partial compliance studies, improving upon the results of Manski. In Chapter 6, we extend these results to the case where the domain of treatment values is continuous, and show how these bounds may be further tightened when the continuous domain is partitioned into ranges of homogeneous treatment responses. In Chapter 7 we discuss the potential application of counterfactual probabilities in legal cases, and we present a hypothetical case where proper treatment of counterfactual probabilities is important for correctly determining liability in product-safety litigation. In Chapter 8 we demonstrate the application of counterfactual reasoning to economic policy-making when knowledge is given by structural equation models.

Part II

Computation

CHAPTER 2

Counterfactual probabilities

2.1 Introduction

This chapter will show that causal theories specified in functional form (as in [PV91, DS93, Poo93]) are sufficient for evaluating counterfactual queries, whereas the causal information embedded in Bayesian networks is not sufficient for the task. Every Bayes network can be represented by several functional specifications, each yielding different evaluations of a counterfactual. The problem is that, deciding what factual information deserves undoing (by the antecedent of the query) requires a model of temporal persistence, and, as noted in [Pea93d], such a model is not part of static Bayesian networks. A functional specification, however, implicitly contains the necessary temporal persistence information.

The next section will introduce some notation for concisely expressing *counterfactual probabilities*. Section 2.3 will describe the relationship between probabilistic and functional specifications, and will demonstrate that probabilistic specifications do not provide sufficient information for precisely evaluating counterfactual probabilities. Section 2.4 will provide an algorithm for evaluating counterfactual probabilities, given a functional model of the system under query. The algorithm will then be applied to the Firing-Squad example introduced in the previous chapter. In Section 2.8 we will describe how counterfactual conditionals may be analyzed when functional assumptions (e.g., linear-normal models) are imposed on a model.

It is assumed that the reader is already familiar with probabilistic causal networks: representation and inference techniques. If not, the reader is referred to [Pea88].

2.2 Notation

Let the set of variables describing the world be designated by $X = \{X_1, X_2, \ldots, X_n\}$. As part of the complete specification of a counterfactual

query, there are real-world observations that make up the background context. These observed values will be represented in the standard form $x_1, x_2, \ldots, x_n$. In addition, we must represent the value of the variables in the counterfactual world. To distinguish between x_i and the value of X_i in the counterfactual world, we will denote the latter with an asterisk; thus, the value of X_i in the counterfactual world will be represented by x_i^*. We will also need a notation to distinguish between events that might be true in the counterfactual world and those referenced explicitly in the counterfactual antecedent. The latter are interpreted as being forced to the counterfactual value by an external intervention, which will be denoted by a hat (e.g., $\hat{x}$).

Thus, a typical counterfactual query will have the form "What is $P(c^*|\hat{a}^*, b)$?" to be read as "Given that we have observed $B = b$ in the real world, if A were a, then what is the probability that C would have been c?"

2.3 Probabilistic vs. functional specification

In this section we will demonstrate that functionally modeled causal theories [PV91] are necessary for uniquely evaluating counterfactual queries, while the conditional probabilities used in the standard specification of Bayesian networks are insufficient for obtaining unique solutions.

Reconsider the firing-squad example limited to the two variables C and B, representing the Captain's signal and Bob's firing, respectively. Assume that previous behavior shows $P(b_1|c_1) = 0.9$ and $P(b_0|c_0) = 0.9$. We observe the Captain give the release signal and Bob not fire, and then wonder with what probability Bob would have fired if the Captain had given the order to fire, i.e., what is $P(b_1^*|\hat{c}_1^*, c_0, b_0)$? The answer depends on the mechanism that accounts for the 10% exception in Bob's behavior. If the reason Bob occasionally does not fire (when the Captain signals to shoot) is that his gun has jammed and he is unable to fire, then the answer to our query would be 8/9 (this result will be evaluated in detail in Section 3.3). However, if the only reason for Bob's occasional non-firing (when the Captain signals to shoot) is that he got the signalling instructions mixed-up, then the answer to our query is 100%, because the Captain's release signal and Bob's non-firing proves that Bob has not mixed up the signals. Thus, we see that the information contained in the conditional probabilities on the observed variables is insufficient for answering counterfactual queries uniquely; some information about the mechanisms responsible for these probabilities is needed as well.

The functional specification, which provides this information, models the influence of C on B by a deterministic function

$$b = F_b(c, \epsilon_b)$$

where ϵ_b stands for all unknown factors that may influence B and the prior probability distribution $P(\epsilon_b)$ quantifies the likelihood of such factors. For example, whether Bob's gun is jammed and whether Bob has the signals crossed could make up two possible components of ϵ_b. Given a specific value for ϵ_b, B becomes a deterministic function of C; hence, each value in ϵ_b's domain specifies a *response function* that maps each value of C to some value in B's domain. In general, the domain for ϵ_b could contain many components, but it can always be replaced by an equivalent variable that is minimal, by partitioning the domain into equivalence regions, each corresponding to a single response function [Pea93a]. Formally, these equivalence classes can be characterized as a function $r_b : \mathrm{dom}(\epsilon_b) \rightarrow \mathbf{N}$, as follows:

$$r_b(\epsilon_b) \;=\; \begin{cases} 0 & \text{if } F_b(c_0, \epsilon_b) = 0 \;\&\; F_b(c_1, \epsilon_b) = 0 \\ 1 & \text{if } F_b(c_0, \epsilon_b) = 0 \;\&\; F_b(c_1, \epsilon_b) = 1 \\ 2 & \text{if } F_b(c_0, \epsilon_b) = 1 \;\&\; F_b(c_1, \epsilon_b) = 0 \\ 3 & \text{if } F_b(c_0, \epsilon_b) = 1 \;\&\; F_b(c_1, \epsilon_b) = 1 \end{cases}$$

Obviously, r_b can be regarded as a random variable that takes on as many values as there are functions between C and B. This domain-minimal variable will be referred to as a *response-function variable.* r_b is closely related to the *potential response variables* in Rubin's model of counterfactuals [Rub74], which was introduced to facilitate causal inference in statistical analysis [BP93].

Suppose that a variable X has causal influences $\{U_1, U_2, \ldots, U_k\}$ in a probabilistic causal model. Let the domain size of each influence U_i be given by m_i, and the domain size of X be given by n. The domain size of X's response-function variable R_x will then be of size

$$n^m \tag{2.1}$$

where

$$m \;=\; \prod_{i=1}^{k} m_i \tag{2.2}$$

This suggests that more than anything else, the domain sizes and fan-in of variables will be the main contributing factors to the computational complexity of evaluating counterfactual probabilities.

For this example, the response-function variable for B has a four-valued domain $r_b \in \{0, 1, 2, 3\}$ with the following functional specification:

$$b = f_b(c, r_b) = h_{b,r_b}(c) \tag{2.3}$$

where the mappings defined by each response function $h_{b,r_b}(c)$ are given by

$$h_{b,0}(c) = b_0 \tag{2.4}$$

$$h_{b,1}(c) = \begin{cases} b_0 & \text{if } c = c_0 \\ b_1 & \text{if } c = c_1 \end{cases} \tag{2.5}$$

$$h_{b,2}(c) = \begin{cases} b_1 & \text{if } c = c_0 \\ b_0 & \text{if } c = c_1 \end{cases} \tag{2.6}$$

$$h_{b,3}(c) = b_1 \tag{2.7}$$

The prior probability of these response functions $P(r_b)$ in conjunction with $f_b(c, r_b)$ fully parameterizes the relationship between C and B in the model.

For each observable variable X_i, there is a function that maps the value of X_i's observable causal influences $\text{pa}(X_i)$ and X_i's response-function variable r_{x_i} to the value of X_i

$$x_i = f_{x_i}(\text{pa}(x_i), r_{x_i})$$

If the model is *complete* (such as the functional model described in [PV91]), all response functions will be mutually independent, and each will be characterized by a prior probability $P(r_{x_i})$. However, when some variables are left out of the analysis, the response functions of the remaining variables $(x_1, \ldots, x_n)$ may be dependent and, in principle, a joint probability $P(r_{x_1}, \ldots, r_{x_n})$ would be required. In practice, only local dependencies will be needed.

If one assumes that two variables C and B are dependent via some exogenous common cause, then we create an edge between r_c and r_b and specify the joint distribution $P(r_c, r_b)$. This treatment of latent variables will be utilized in the applications discussed in Sections 5.1 and 7.2.

Given $P(r_b)$, we can uniquely evaluate the counterfactual query "What is $P(b_1^*|\hat{c}_1^*, c_0, b_0)$?" (i.e., "Given $C = c_0$ and $B = b_0$, if C were c_1, then what is the probability that B would have been b_1?"). The intervention-based interpretation of counterfactual antecedents implies that the disturbance ϵ_b, and hence the response-function r_b, is unaffected by the interventions that force the counterfactual values; therefore, what we learn about the response-function from the observed evidence is applicable to the evaluation of belief in the counterfactual con-

sequent. If we observe (c_0, b_0), then we are certain that $r_b \in \{0,1\}$, an event having prior probability $P(r_b{=}0)+P(r_b{=}1)$. Hence, this evidence leads to an updated posterior probability for r_b (let $\vec{P}(r_b) = \langle P(r_b{=}0), P(r_b{=}1), P(r_b{=}2), P(r_b{=}3)\rangle$)

$$\vec{P}'(r_b) = \vec{P}(r_b|c_0, b_0) = \langle \frac{P(r_b{=}0)}{P(r_b{=}0)+P(r_b{=}1)}, \frac{P(r_b{=}1)}{P(r_b{=}0)+P(r_b{=}1)}, 0, 0\rangle.$$

According to Eqs. 2.3-2.7, if C were forced to c_1, then B would have been b_1 if and only if $r_b \in \{1,3\}$, which has probability $P'(r_b{=}1) + P'(r_b{=}3) = P'(r_b{=}1)$. This is exactly the solution to the counterfactual query,

$$P(b_1^*|\hat{c}_1^*, c_0, b_0) = P'(r_b{=}1) = \frac{P(r_b{=}1)}{P(r_b{=}0)+P(r_b{=}1)}.$$

This analysis is consistent with the *prior propensity account* of [Sky80].

What if we are provided only with the conditional probability $P(b|c)$ instead of the functional model ($f_b(c, r_b)$ and $P(r_b)$)? These two specifications are related by:

$$\begin{aligned} P(b_1|c_0) &= P(r_b{=}2) + P(r_b{=}3) \\ P(b_1|c_1) &= P(r_b{=}1) + P(r_b{=}3). \end{aligned}$$

which show that $P(r_b)$ is not, in general, uniquely determined by the conditional distribution $P(b|c)$.

Hence, given a counterfactual query, a functional model always leads to a unique solution, while a Bayesian network seldom leads to a unique solution, depending on whether the conditional distributions of the Bayesian network sufficiently constrain the prior distributions of the response-function variables in the corresponding functional model. In Chapter 3 we will develop techniques for evaluating bounds on counterfactual probabilities when only given conditional probability distributions on the observable variables.

In practice, specifying a functional model is not as daunting as one might think from the example above. In fact, it could be argued that the subjective judgments needed for specifying Bayesian networks (i.e., judgments about conditional probabilities) are generated mentally on the basis of a stored model of functional relationships. For example, in the noisy-OR mechanism, which is often used to model causal interactions, the conditional probabilities are derivatives of a functional model involving AND/OR gates, corrupted by independent binary disturbances. This model is used, in fact, to *simplify* the specification of conditional probabilities in Bayesian networks [Pea88].

2.4 Evaluating counterfactual queries

From the last section, we see that the algorithm for evaluating counterfactual queries should consist of the following steps: (1) compute the posterior probabilities for the disturbance variables, given the observed evidence; (2) remove the observed evidence and enforce the value for the counterfactual antecedent; finally, (3) evaluate the probability of the counterfactual consequent, given the conditions set in the first two steps.

An important point to remember is that it is not enough to compute the posterior distribution of each disturbance variable (ϵ) separately and treat those variables as independent quantities. Although the disturbance variables are initially independent, the evidence observed tends to create dependencies among the parents of the observed variables, and these dependencies need to be represented in the posterior distribution. An efficient way to maintain these dependencies is through the structure of the causal network itself.

Thus, we will represent the variables in the counterfactual world as distinct from the corresponding variables in the real world, by using a separate network for each world. Evidence can then be instantiated on the real-world network, and the solution to the counterfactual query can be determined as the probability of the counterfactual consequent, as computed in the counterfactual network where the counterfactual antecedent is enforced. But, the reader may ask, and this is key, how are the networks for the real and counterfactual worlds linked? Because any exogenous variable, ϵ_a, is not influenced by forcing the value of any endogenous variables in the model, the value of that disturbance will be identical in both the real and counterfactual worlds; therefore, a single variable can represent the disturbance in both worlds. ϵ_a thus becomes a common causal influence of the variables representing A in the real and counterfactual networks, respectively, which allows evidence in the real-world network to propagate to the counterfactual network.

Assume that we are given a *causal theory* $T = \langle D, \Theta_D \rangle$ as defined in [PV91]. D is a directed acyclic graph (DAG) that specifies the structure of causal influences over a set of variables $X = \{X_1, X_2, \ldots, X_n\}$. Θ_D specifies a functional mapping $x_i = f_i(\text{pa}(x_i), \epsilon_i)$ ($\text{pa}(x_i)$ represents the value of X_i's parents) and a prior probability distribution $P(\epsilon_i)$ for each disturbance ϵ_i (we assume that ϵ_i's domain is discrete; if not, we can always transform it to a discrete domain such as a response-function variable). A counterfactual query "What is $P(c^*|\hat{a}^*, o)$?" is then posed, where c^* specifies counterfactual values for a set of variables $C \subset X$,

$\hat{a}^*$ specifies forced values for the set of variables in the counterfactual antecedent, and o specifies observed evidence. The solution can be evaluated by the following algorithm:

1. From the known causal theory T create a Bayesian network $< G, \mathcal{P} >$ that explicitly models the disturbances as variables and distinguishes the real world variables from their counterparts in the counterfactual world. G is a DAG defined over the set of variables $V = X \cup X^* \cup \epsilon$, where $X = \{X_1, X_2, \ldots, X_n\}$ is the original set of variables modeled by T, $X^* = \{X_1^*, X_2^*, \ldots, X_n^*\}$ is their counterfactual world representation, and $\epsilon = \{\epsilon_1, \epsilon_2, \ldots, \epsilon_n\}$ represents the set of disturbance variables that summarize the common external causal influences acting on the members of X and X^*. $\mathcal{P}$ is the set of conditional probability distributions $P(V_i|\mathrm{pa}(V_i))$ that parameterizes the causal structure G.

   If $X_j \in \mathrm{pa}(X_i)$ in D, then $X_j \in \mathrm{pa}(X_i)$ and $X_j^* \in \mathrm{pa}(X_i^*)$ in G ($\mathrm{pa}(X_i)$ is the set of X_i's parents). In addition, $\epsilon_i \in \mathrm{pa}(X_i)$ and $\epsilon_i \in \mathrm{pa}(X_i^*)$ in G. The conditional probability distributions for the Bayesian network are generated from the causal theory:

$$P(x_i|\mathrm{pa}_X(x_i), \epsilon_i) = \begin{cases} 1 & \text{if } x_i = f_i(\mathrm{pa}_X(x_i), \epsilon_i) \\ 0 & \text{otherwise} \end{cases}$$

   where $\mathrm{pa}_X(x_i)$ is the set of values of the variables in $X \cap \mathrm{pa}(x_i)$.

$$P(x_i^*|\mathrm{pa}_{X^*}(x_i^*), \epsilon_i) = P(x_i|\mathrm{pa}_X(x_i), \epsilon_i)$$

   whenever $x_i = x_i^*$ and $\mathrm{pa}_{X^*}(x_i^*) = \mathrm{pa}_X(x_i)$. $P(\epsilon_i)$ is the same as specified by the functional causal theory T.

2. Observed evidence. The observed evidence o is instantiated on the real world variables X corresponding to o.

3. Counterfactual antecedent. For every forced value in the counterfactual antecedent specification $\hat{x}_i^* \in \hat{a}^*$, apply the intervention-based semantics of $set(X_i^* = \hat{x}_i^*)$ (see [Pea93a, SGS93]), which amounts to severing all the causal edges from $\mathrm{pa}(X_i^*)$ to X_i^* for all $x_i^* \in \hat{a}^*$ and instantiating X_i^* to the value specified in $\hat{a}^*$.

4. Belief propagation. After instantiating the observations and interventionss in the network, evaluate the belief in c^* using the standard belief update methods for Bayesian networks [Pea88]. The result is the solution to the counterfactual query.

Note that all evidence does not have to come in the form of concrete observations; evidence can also be given as likelihood information. For example, we may receive a report from one of the other guards that he saw a pardon on the Warden's desk, but he did not know which Traitor it was for. We may quantify this evidence ($e \equiv$ guard seeing pardon) by a likelihood vector which indicates the relative chance that the evidence came in given the Captain's signal, i.e., $\lambda_C = \langle P(e|c_0), P(e|c_1) \rangle$. [Pea88] describes how such evidence is used to update our beliefs in a Bayesian network. Additional notation is needed to add this virtual evidence into the specification of a counterfactual probability.

In the last section, we noted that the conditional distribution $P(x_k|\text{pa}(X_k))$ for each variable $X_k \in X$ constrains, but does not uniquely determine, the prior distribution $P(\epsilon_k)$ of each disturbance variable. Although the composition of the external causal influences are often not precisely known, a subjective distribution over response functions may be assessable. If a reasonable distribution can be selected for each relevant disturbance variable, the implementation of the above algorithm is straightforward and the solution is unique; otherwise, bounds on the solution can be obtained using convex optimization techniques. In the next chapter, we will explain how such optimization tasks are formulated, and Chapter 5 applies this technique for deriving bounds on causal effects from partially controlled experiments.

2.5 Firing Squad Revisited

Let us revisit the firing squad example. Assuming we have observed that Bob fired his rifle ($b = b_1$), we want to know with what probability the Traitor would have lived if Bob had not fired his rifle (i.e., "What is $P(t_1^*|\hat{b}_0^*, b_1)$?").

Suppose that we are supplied with the following causal theory for the model in Figure 1.1:

$$\begin{aligned} c = &\quad f_c(r_c) &&= h_{c,r_c}() \\ b = &\quad f_b(c, r_b) &&= h_{b,r_b}(c) \\ t = &\quad f_t(b, c, r_t) &&= h_{t,r_t}(b, c) \end{aligned}$$

where

$$P(r_c) = \begin{cases} 0.40 & \text{if } r_c = 0 \\ 0.60 & \text{if } r_c = 1 \end{cases}$$

$$P(r_b) = \begin{cases} 0.02 & \text{if } r_b = 0 \\ 0.90 & \text{if } r_b = 1 \\ 0.08 & \text{if } r_b = 2 \\ 0 & \text{if } r_b = 3 \end{cases}$$

$$P(r_t) = \begin{cases} 0.01 & \text{if } r_t = 0 \\ 0.40 & \text{if } r_t = 8 \\ 0.09 & \text{if } r_t = 10 \\ 0.35 & \text{if } r_t = 12 \\ 0.13 & \text{if } r_t = 14 \\ 0.02 & \text{if } r_t = 15 \\ 0 & \text{otherwise} \end{cases}$$

and

$$h_{c,0}() = c_0$$
$$h_{c,1}() = c_1$$

$$h_{t,0}(b,c) = t_0$$

$$h_{t,1}(b,c) = \begin{cases} t_0 & \text{if } (b,c) \neq (b_1,c_1) \\ t_1 & \text{if } (b,c) = (b_1,c_1) \end{cases}$$

$$h_{t,2}(b,c) = \begin{cases} t_0 & \text{if } (b,c) \neq (b_0,c_1) \\ t_1 & \text{if } (b,c) = (b_0,c_1) \end{cases}$$

$$h_{t,3}(b,c) = \begin{cases} t_0 & \text{if } c = c_0 \\ t_1 & \text{if } c = c_1 \end{cases}$$

$$h_{t,4}(b,c) = \begin{cases} t_0 & \text{if } (b,c) \neq (b_1,c_0) \\ t_1 & \text{if } (b,c) = (b_1,c_0) \end{cases}$$

$$h_{t,5}(b,c) = \begin{cases} t_0 & \text{if } b = b_0 \\ t_1 & \text{if } b = b_1 \end{cases}$$

$$h_{t,6}(b,c) = \begin{cases} t_0 & \text{if } (b,c) \in \{(b_0,c_0),(b_1,c_1)\} \\ t_1 & \text{if } (b,c) \in \{(b_1,c_0),(b_0,c_1)\} \end{cases}$$

$$h_{t,7}(b,c) = \begin{cases} t_0 & \text{if } (b,c) = (b_0,c_0) \\ t_1 & \text{if } (b,c) \neq (b_0,c_0) \end{cases}$$

$$h_{t,8}(b,c) = \begin{cases} t_0 & \text{if } (b,c) \neq (b_0,c_0) \\ t_1 & \text{if } (b,c) = (b_0,c_0) \end{cases}$$

$$h_{t,9}(b,c) = \begin{cases} t_0 & \text{if } (b,c) \in \{(b_1,c_0),(b_0,c_1)\} \\ t_1 & \text{if } (b,c) \in \{(b_0,c_0),(b_1,c_1)\} \end{cases}$$

$$
\begin{aligned}
h_{t,10}(b,c) &= \begin{cases} t_0 & \text{if } b = b_1 \\ t_1 & \text{if } b = b_0 \end{cases} \\
h_{t,11}(b,c) &= \begin{cases} t_0 & \text{if } (b,c) = (b_1,c_0) \\ t_1 & \text{if } (b,c) \neq (b_1,c_0) \end{cases} \\
h_{t,12}(b,c) &= \begin{cases} t_0 & \text{if } c = c_1 \\ t_1 & \text{if } c = c_0 \end{cases} \\
h_{t,13}(b,c) &= \begin{cases} t_0 & \text{if } (b,c) = (b_0,c_1) \\ t_1 & \text{if } (b,c) \neq (b_0,c_1) \end{cases} \\
h_{t,14}(b,c) &= \begin{cases} t_0 & \text{if } (b,c) = (b_1,c_1) \\ t_1 & \text{if } (b,c) \neq (b_1,c_1) \end{cases} \\
h_{t,15}(b,c) &= t_1
\end{aligned}
$$

The response functions for B (h_{b,r_b}) take the same form as that given in Eq. (2.7).

These numbers reflect the authors' understanding of the dynamics involved. For example, the choice for $P(r_b)$ represents our belief that Bob usually fires if and only if the Captain gives the signal to fire. However, we believe that Bob is sometimes ($\sim 2\%$ of the time) unable to fire (e.g., his gun jams); this exception is represented by $r_b = 0$. In addition, Bob sometimes ($\sim 3\%$ of the time) fires if and only if the Captain gives the order to release the Traitor (e.g., Bob has his signals crossed); this exception is represented by $r_b = 2$.

Finally, $P(r_t)$ represents our understanding that there is a slight chance (1%) that the Traitor is "scared to death" ($r_t = 0$) and a slight chance (2%) that all of the riflemen miss their target ($r_t = 15$). In addition, the chances that different combinations of riflemen inflict a lethal wound are broken down as follows: 40% of the time both Bob and the other riflemen are on their mark ($r_t = 8$); 9% of the time only Bob is on his mark ($r_t = 10$); 35% of the time only the other riflemen are on their mark ($r_t = 12$); and 13% of the time it takes the combined influence of Bob and the other riflemen to inflict a lethal wound ($r_t = 14$).

Figure 2.1 shows the Bayesian network generated from step 1 of the algorithm. After instantiating the real world observations (b_0) and the interventions ($\hat{b}_1^*$) specified by the counterfactual antecedent in accordance with steps 2 and 3, the network takes on the configuration shown in Figure 2.2.

If we propagate the evidence through this Bayesian network, we will arrive at the solution

$$P(t_1^*|\hat{b}_0^*, b_1) = 0.15$$

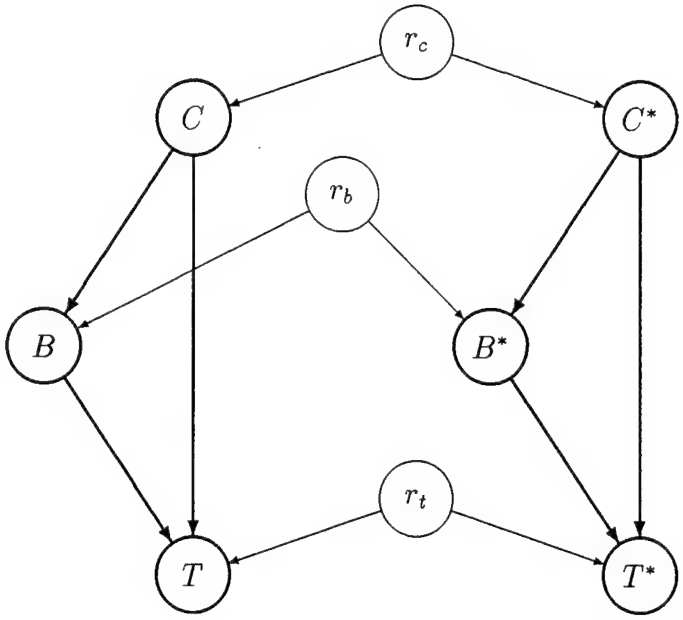

Figure 2.1: *Bayesian model for evaluating counterfactual queries in the firing-squad example. The variables marked with $*$ make up the counterfactual world, while those without $*$, the factual world. The r variables index the response functions.*

which is consistent with Dave's assertion that the Traitor would still have died had Bob not fired, given that Bob had actually fired. Compare this with the solution to Scott's indicative counterfactual query:

$$P(t_1|b_0) \;=\; 0.88.$$

that is, if we had observed that Bob did not fire, the Traitor probably would not have died. This emphasizes the difference between the intervention-based interpretation and a revisionist interpretation of counterfactual conditionals.

2.6 Complexity Issues

The complexity of belief update in a probabilistic network is dependent on the structure of the network along with the variables for which evidence is available. If the structure of causal knowledge is given by a directed tree, then belief update occurs in parallel at each node, requiring only a polynomial number of calculations in terms of the variables' domain sizes. If the network is not a directed tree, but there are no loops in the network (i.e., polytrees), then the complexity becomes

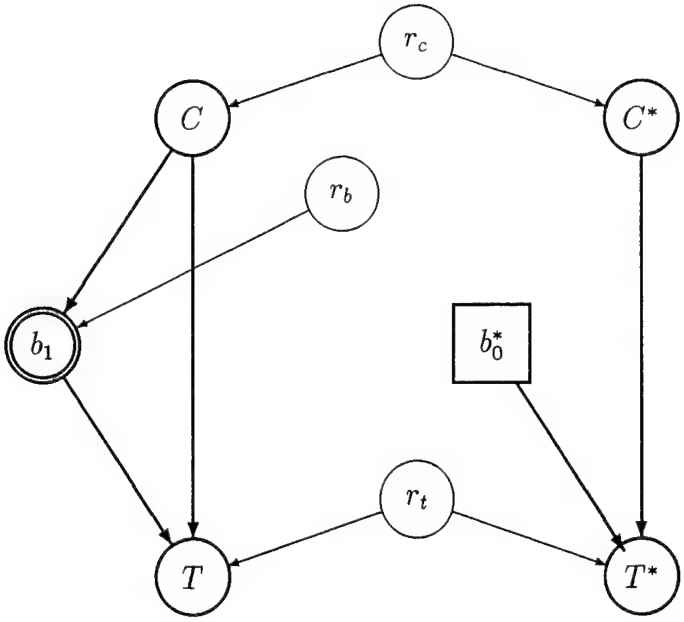

Figure 2.2: *To evaluate the query $P(t_1^*|\hat{b}_0^*, b_1)$, the network of Figure 2.1 is instantiated with observation b_1 and intervention $\hat{b}_0^*$ (links pointing to b_0^* are severed).*

exponential in terms of the number of parents of each variable [Pea88, p. 183]. For unrestricted directed acyclic graphs, the greatest increase in complexity comes about from cycles in the structure induced by observations on child variables. For example, if the causal structure is given by $A \rightarrow B$, $A \rightarrow C$, $B \rightarrow D$, and $C \rightarrow D$, then a cycle would be induced on the network by observation of D. [Pea88] discusses various ways of updating beliefs when induced cycles are present in the network.

These same results apply to the computation of counterfactual probabilities; however, given a probabilistic causal model for a system, computing a counterfactual probability is much more expensive than computing a similar conditional probability, because response-function variables must be specified whose domains grow in size according to Eqs. (2.1) and (2.2).

A network generated by the algorithm in Section 2.4 may often be simplified, because not all response-function variables need to be generated, as dictated by the following theorem:

Theorem 2.6.1 *A response-function variable r_x for the variable X is necessary for evaluating a counterfactual probability $P(c^*|\hat{a}^*, o)$ if and only if*

- *X is a descendant of any of the variables specified in the counterfactual*

antecedent $\hat{a}^*$,

- *evidence is available for either* X *or one of its descendants in the factual world, and*
- *the relationship between* X *and its known causal influences is nondeterministic.*

Proof:

If a variable X_j^* in the counterfactual world is not a causal descendant of any of the variables mentioned in the counterfactual antecedent $\hat{a}^*$, then X_j and X_j^* will always have identical distributions, because the causal influences that functionally determine X_j and X_j^* are identical. X_j and X_j^* may therefore be treated as the same variable. In this case, the conditional distribution $P(x_j|\text{pa}(x_j))$ is sufficient, and the disturbance variable ϵ_j and its prior distribution need not be specified.

If X_j^* is a causal descendant of one of the variables in the counterfactual antecedent, but neither X_j nor X_j's descendants have been observed in the real world, then the observations in the real world provide no information about the distribution of X_j's response-functions r_{x_j}. Therefore, all we need to know is the conditional probability distribution $P(x_j|\text{pa}(x_j))$ for evaluating the counterfactual probability.

If X is already a deterministic function of its causal influences, then a response-function variable becomes redundant, because one of the response functions will have probability of one, and no observed evidence will change this prior distribution. Thus, the functional mapping remains the same in both the real and counterfactual worlds.

However, if all three conditions above are true, then a response-function variable r_x is necessary because (1) the evidence on X and/or propagated from its descendants produces a posterior distribution on r_x, that, in essence, changes the conditional distribution $P(x_j|\text{pa}(x_j))$ in the counterfactual world; and (2) the causal influences on X (besides r_x) have different values (or a different distribution of values) between the real and counterfactual worlds. This means that we must know how the mapping from one valuation of causal influences to a child value is related to the mapping from another valuation of the causal influences; this relationship is provided by the distribution on the response-function variable.

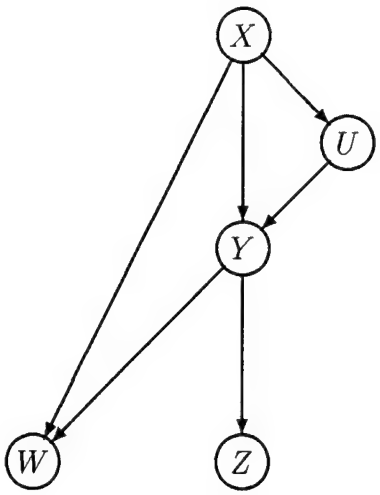

Figure 2.3: *Hypothetical model with full functional specification (all response-function variables).*

□

Important in this discussion is that for evaluating a particular counterfactual probability a specification of response-functions and their prior distribution are only necessary for a subset of the variables in the probabilistic causal model. Consider a causal model over the variables $\{U, W, X, Y, Z\}$ with the structure shown in Figure 2.3. This model is parameterized by the the conditional probability distributions $P(x)$, $P(u|x)$, $P(y|x,u)$, $P(w|x,y)$, and $P(z|y)$ and consists of

$$(M_x - 1) + M_x(M_u - 1) + M_x M_u(M_y - 1) + M_x M_y(M_w - 1) + M_y(M_z - 1)$$

independent parameters, where M_x is the variable X's domain size.

In order to evaluate counterfactual probabilities for this model in general, we would generate the combined functional model for the factual and counterfactual worlds. The structure for this functional model is shown in Figure 2.4. This structure is parameterized by the prior probability distributions on the response-function variables, which consists of

$$(M_x - 1) + (M_u^{M_x} - 1) + (M_y^{M_x M_u} - 1) + (M_w^{M_x M_y} - 1) + (M_z^{M_y} - 1)$$

independent parameters. It would be desirable to avoid having to specify all parameters associated with the response-function distributions. Fortunately, not

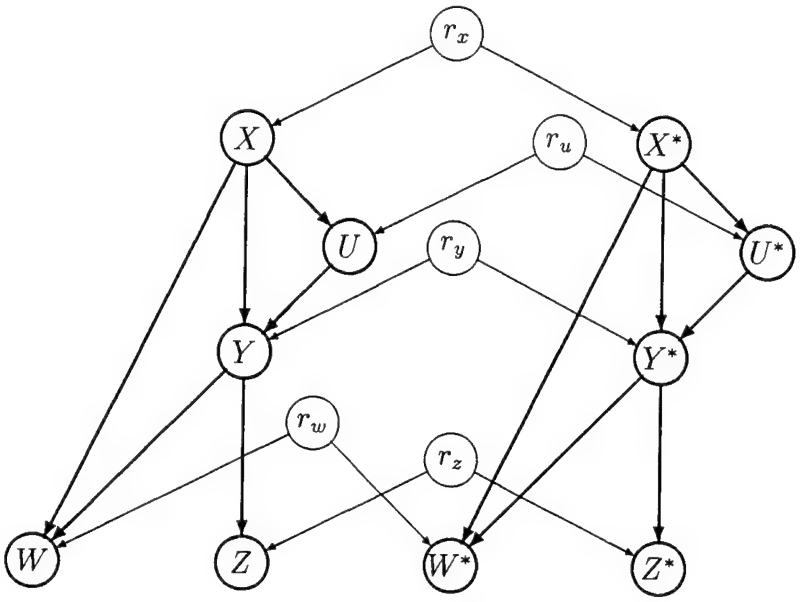

Figure 2.4: *Hypothetical model with full functional specification (all response-function variables).*

all response-functions are relevant to the evaluation of specific counterfactual probabilities. For example, suppose that we need to evaluate $P(x^*, y^*, z^*|\hat{u}^*, w)$.

Applying Theorem 2.6.1 we can eliminate from consideration the response-function variables for X, U, and Z. r_x is eliminated, because X's causal influences in the factual and counterfactual world will always be the same; therefore, we may treat X and X^* as one and the same (we will notate this variable by $X^{(*)}$), and only need to specify $P(x)$. r_u is eliminated, because U^* is subjected to a local change despite the value of the response-function; therefore, no information is conveyed from U to U^* and we only need to specify $P(u|x)$. r_z is eliminated, because the observations in the factual world do not propagate to r_z; therefore, the posterior probability on r_z is identical to its prior probability. Hence, $P(z|y)$ is sufficient for parameterizing the relation between Y^* and Z^* in the counterfactual world. In fact, the variable Z in the real world is irrelevant to the evaluation of the counterfactual probability and may itself be eliminated. If we perform these simplifications in the model, we are left with the structure shown in Figure 2.5.

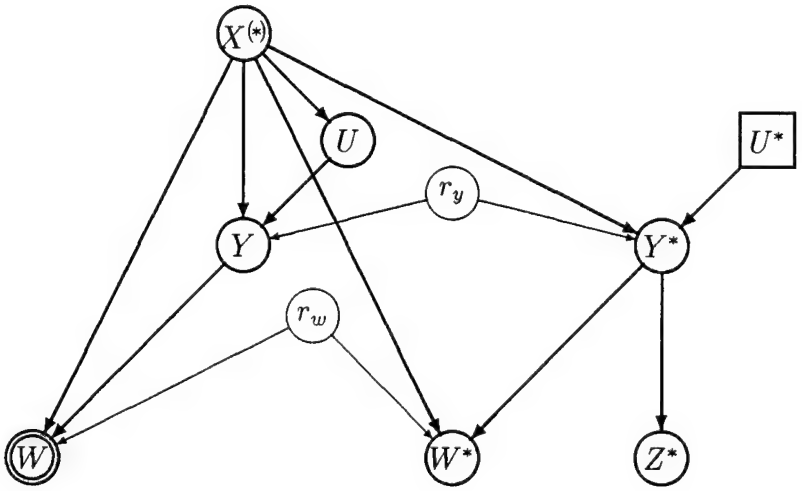

Figure 2.5: *Simplified functional specification for a given observation on W and counterfactual antecedent specifying just U. Note that the response functions r_x, r_u, and r_z do not require specification.*

2.7 Statistical independence and counterfactual probabilities

Statistical independence, e.g., $P(b_1|a_0) = P(b_1|a_1)$, does not give us permission to remove a causal edge from a probabilistic specification. If there is in fact a direct causal influence from A to B, then a functional specification for the model can lead to divergent values for the counterfactual probability $P(b_1^*|\hat{a}_1^*, a_0, b_0)$. For example, suppose that

$$\begin{aligned} P(b_1|a_0) &= 0.50 \\ P(b_1|a_1) &= 0.50 \end{aligned}$$

We can imagine two distributions over B's response functions consistent with the conditional probability distribution $P(b|a)$:

$$\begin{aligned} P_1(r_b{=}0) &= 0.5 \\ P_1(r_b{=}1) &= 0.0 \\ P_1(r_b{=}2) &= 0.0 \\ P_1(r_b{=}3) &= 0.5 \end{aligned}$$

and

$$\begin{aligned} P_2(r_b{=}0) &= 0.0 \\ P_2(r_b{=}1) &= 0.5 \\ P_2(r_b{=}2) &= 0.5 \\ P_2(r_b{=}3) &= 0.0 \end{aligned}$$

The first distribution $P_1(b|a)$ does verify the independence of A and B and evaluates $P(b_1^*|\hat{a}_1^*, a_0, b_0) = 0.0$, while the second distribution $P_2(b|a)$ shows B deterministically dependent on A and evaluates $P(b_1^*|\hat{a}_1^*, a_0, b_0) = 1.0$. Thus, it is crucial that the dependencies in the causal model are determined by more than statistical considerations, but also by subjective knowledge of causal effects.

2.8 Parametric and canonical models

The advantage of using specialized models, e.g., parametric or canonical, arises from the reduction in the number of parameters necessary for completely specifying the model. The exponential size of response-function variable domains

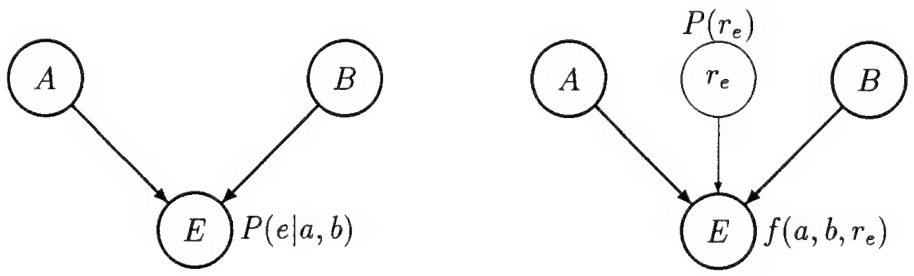

Figure 2.6: *Unconstrained model of two known variables influencing a third.*

provides a strong incentive to develop and apply these canonical representations. The reduction in parameters can possibly lead to more efficient computations, but will never lead to less efficiency, because the complete response-function model may always be generated from the reduced-parameter canonical model.

2.8.1 Canonical models

The typical method for reducing the number of parameters in probabilistic causal networks is to decompose the relationship between an effect and its set of causes into an expanded model with additional variables that impose structural independencies. For example, suppose that two binary variables A and B causally influence another binary variable E as depicted in Figure 2.6. In an unrestricted model, the conditional probability distributions $P(E|a,b)$, $a \in \{a_0, a_1\}$ and $b \in \{b_0, b_1\}$, require the specification of four independent parameters, and the distribution for the response-function variable r_e, $P(r_e)$, requires the specification of 15 independent parameters.

Suppose, however, that the interaction between these variables is correctly modelled by a Noisy-OR gate. This model imposes additional structural assumptions into the causal network, as depicted in Figure 2.7. Two new intermediate variables, I_a and I_b, have been introduced into the network; E is functionally determined by I_a and I_b ($E = I_a \vee I_b$), and the nondeterministic effects of A on I_a, and B on I_b are specified by the conditional probability distributions $P(I_a|A)$ and $P(I_b|B)$. This Noisy-OR model still requires the specification of four independent parameters in the general case, but the savings become apparent when attempting to evaluate counterfactual probabilities. From Theorem 2.6.1, no

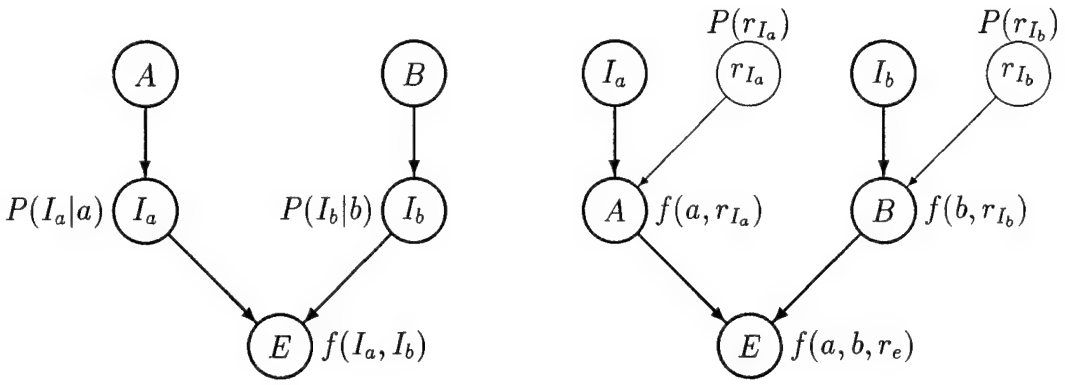

Figure 2.7: *Functional model assuming that the influence of A and B on E may be modelled by a Noisy-OR gate.*

response-function variable need be generated for E, because E is a deterministic function of its causal parents. Instead, response-function variables are generated for I_a and I_b, and the distribution of the response functions, $P(r_{I_a})$ and $P(r_{I_b})$ are specified; however, this requires specification of only six independent parameters.

The reduction in parameters becomes even more drastic as the number of causal influences impinging on a variable increase. Consider the case where n binary variables C_1, C_2, ..., C_n influence another binary variable E. The conditional probability distributions $P(e|c_1, c_2, \ldots, c_n)$ are completely specified by $2^n - 1$ independent parameters. The general functional-model for this pattern of influence is depicted in Figure 2.8 where the distribution of response-functions $P(r_e)$ requires specification of $2^{(2^n)} - 1$ independent parameters. This super-exponential growth of parameters as a function of the number of causal influences makes the task of counterfactual inference unmanageable. It quickly becomes apparent that the number of specification parameters must be reduced in order to make any headway.

Suppose we assume that the relationship between a variable and its causal influences satisfies the temporal definition of causal independence [Hec93]. In this case, the causal structure of Figure 2.8 may be expanded to the structure depicted in Figure 2.9. Each response-function variable $r_{e_1}, \ldots, r_{e_n}$ specifies the mapping from two binary variables to a single binary variable (requiring specification of 15 independent parameters for $P(r_{e_k})$), while the prior distribution on r_{e_0} is just the same as the prior distribution on e_0. Therefore, the decomposed model of

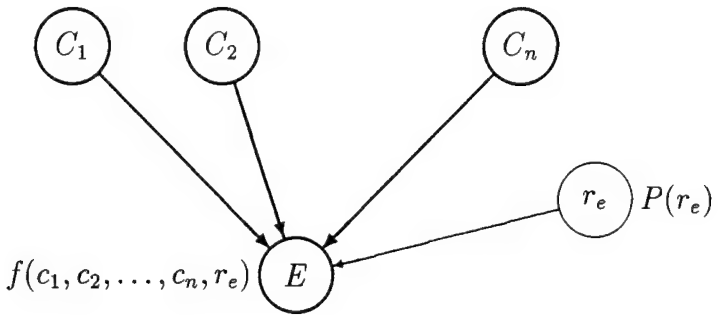

Figure 2.8: *Unconstrained model with n known variables influencing the variable E.*

Figure 2.9 requires the specification of only $15n + 1$ independent parameters, a drastic reduction from the $2^{(2^n)} - 1$ required for the unconstrained functional model of Figure 2.8.

Of course, a decomposed model should only be applied when it is believed that it provides a good approximation of the relationships existing in the real world. However, from the comparison of parameter counts, it is apparent that the full response-function distribution $P(r_e)$ could be pragmatically unmanageable, and require the use of a decomposed model in order to make any progress.

2.8.2 Linear-Normal Models

Assume that knowledge is specified by the structural equation model (often used in econometrics and the social sciences, and originally established by Sewall Wright in his development of path analysis models [Wri21])

$$\vec{x} = B\vec{x} + \vec{\epsilon}$$

where B is a matrix (not necessarily triangular) corresponding to a causal model (possibly cyclic), and we are given the mean $\vec{\mu}_\epsilon$ and covariance $\Sigma_{\epsilon,\epsilon}$ of the disturbances $\vec{\epsilon}$ (assumed to be normal). The variables on the right-hand side of a structural equation are interpreted as the causal influences of the variable on the left-hand side of the equation. The mean and covariance of the observable variables $\vec{X}$ are then given by:

$$\vec{\mu}_x = S\vec{\mu}_\epsilon \tag{2.8}$$

$$\Sigma_{x,x} = S\Sigma_{\epsilon,\epsilon}S^t \tag{2.9}$$

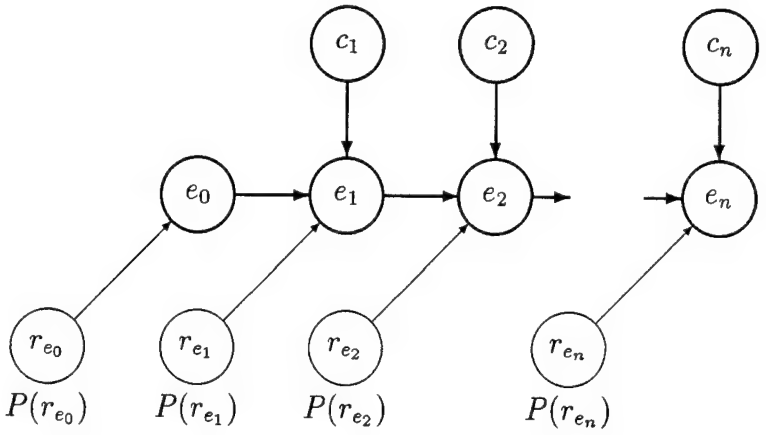

Figure 2.9: *Canonical model assuming temporal causal independence.*

where $S = (I - B)^{-1}$.

Under such a model, there are well-known formulas [Whi90, p. 163] [Dem69] for evaluating the mean and covariance of $\vec{X}$ conditioned on some observations $\vec{o}$:

$$\vec{\mu}_{x|o} = \vec{\mu}_x + \Sigma_{x,o}\Sigma_{o,o}^{-1}(\vec{o} - \vec{\mu}_o) \tag{2.10}$$

$$\Sigma_{x,x|o} = \Sigma_{x,x} - \Sigma_{x,o}\Sigma_{o,o}^{-1}\Sigma_{o,x} \tag{2.11}$$

where, for every pair of sub-vectors, $\vec{Z}$ and $\vec{W}$, of $\vec{X}$, $\Sigma_{z,w}$ is the sub-matrix of $\Sigma_{x,x}$ with entries corresponding to the components of $\vec{Z}$ and $\vec{W}$. Singularities of Σ terms are handled by appropriate means.

Similar formulas apply for the mean and covariance of $\vec{X}$ under an intervention $\vec{a}$. For mathematical convenience, let $\vec{X}$ be partitioned according to whether each variable is referred to in $\vec{a}$. The set of variables referred to in $\vec{a}$ is denoted by $\vec{Z}$, and the set of remaining variables in $\vec{X}$ is denoted by $\vec{Y}$. Under this partition, the matrix B can be partitioned into four submatrices

$$B = \begin{bmatrix} B_{yy} & B_{yz} \\ B_{zy} & B_{zz} \end{bmatrix}$$

B is replaced by the intervention-pruned matrix $\hat{B} = [\hat{b}_{ij}]$ defined by:

$$\hat{b}_{ij} = \begin{cases} 0 & \text{if } X_i \in \vec{a} \\ b_{ij} & \text{otherwise} \end{cases}$$

Equivalently,

$$\hat{B} = \begin{bmatrix} B_{yy} & B_{yz} \\ 0 & 0 \end{bmatrix}$$

According to intervention semantics [Pea94a] all links from $\vec{\epsilon}_z$ to $\vec{Z}$ are severed and $\vec{Z}$ is forced to the value $\vec{a}$. Therefore, the modified structural equation model for $\vec{X}$ when influenced by external interventions is given by

$$\vec{x} = (I - \hat{B})^{-1} \begin{bmatrix} \vec{\epsilon}_y \\ 0 \end{bmatrix} + (I - \hat{B})^{-1} \begin{bmatrix} 0 \\ \vec{a} \end{bmatrix}$$

Given the mean and covariance of $\vec{\epsilon}_y$, the mean and covariance of the observable variables $\vec{X}$ may be evaluated

$$\begin{aligned} \vec{\mu}_{x|\hat{a}} &= \begin{bmatrix} \vec{\mu}_{y|\hat{a}} \\ \vec{\mu}_{z|\hat{a}} \end{bmatrix} \\ &= \begin{bmatrix} (I - B_{yy})^{-1}(\vec{\mu}_{\epsilon_y} + B_{yz}\vec{a}_z) \\ \vec{a}_z \end{bmatrix} \qquad (2.12) \\ \Sigma_{x,x|\hat{a}} &= \Sigma_{yz,yz|\hat{a}} \\ &= \begin{bmatrix} \Sigma_{y,y|\hat{a}} & \Sigma_{y,z|\hat{a}} \\ \Sigma_{z,y|\hat{a}} & \Sigma_{z,z|\hat{a}} \end{bmatrix} \\ &= \begin{bmatrix} (I - B_{yy})^{-1}\Sigma_{\epsilon_y,\epsilon_y}((I - B_{yy})^{-1})^t & 0 \\ 0 & 0 \end{bmatrix} \qquad (2.13) \end{aligned}$$

To evaluate the counterfactual distribution $\mu_{x^*|\hat{a}^*o}$ and $\Sigma_{x^*,x^*|\hat{a}}$ we first update the prior distribution of the disturbances by their distribution conditioned on the observations $\vec{o}$:

$$\begin{aligned} \vec{\mu}^o_\epsilon \triangleq \vec{\mu}_{\epsilon|o} &= \vec{\mu}_\epsilon + \Sigma_{\epsilon,o}\Sigma^{-1}_{o,o}(\vec{o} - \vec{\mu}_o) \\ &= \vec{\mu}_\epsilon + \Sigma_{\epsilon,\epsilon}S_o^t(S_o\Sigma_{\epsilon,\epsilon}S_o^t)^{-1}(\vec{o} - \vec{\mu}_o) \\ \Sigma^o_{\epsilon,\epsilon} \triangleq \Sigma_{\epsilon,\epsilon|o} &= \Sigma_{\epsilon,\epsilon} - \Sigma_{\epsilon,o}\Sigma^{-1}_{o,o}\Sigma_{o,\epsilon} \\ &= \Sigma_{\epsilon,\epsilon} - \Sigma_{\epsilon,\epsilon}S_o^t(S_o\Sigma_{\epsilon,\epsilon}S_o^t)^{-1}S_o\Sigma_{\epsilon,\epsilon} \end{aligned}$$

where S_o is the submatrix of S containing all columns of S, but only those rows corresponding to the observed variables in $\vec{o}$.

We then evaluate the means $\vec{\mu}_{x^*|\hat{a}^*o}$ and variances $\Sigma_{x^*,x^*|\hat{a}^*o}$ of the variables in the counterfactual world ($\vec{X}^*$) under the intervention $\vec{a}$ using Eqs. (2.12) and

(2.13), by replacing the prior distribution on the disturbances $\Sigma_{\epsilon_y,\epsilon_y}$ and μ_{ϵ_y} with the posterior distribution $\Sigma^o_{\epsilon_y,\epsilon_y}$ and $\mu^o_{\epsilon_y}$:

$$\mu_{x^*|\hat{a}^*o} = \begin{bmatrix} (I - B_{yy})^{-1}(\vec{\mu}^o_{\epsilon_y} + B_{yz}a_z) \\ \vec{a}_z \end{bmatrix} \tag{2.14}$$

$$\Sigma_{x^*,x^*|\hat{a}} = \begin{bmatrix} (I - B_{yy})^{-1}\Sigma^o_{\epsilon_y,\epsilon_y}((I - B_{yy})^{-1})^t & 0 \\ 0 & 0 \end{bmatrix} \tag{2.15}$$

It is clear that this procedure can be applied to non-triangular matrices, as long as S is non-singular. An application of this class of model representation will be presented in Chapter 8.

2.9 Conclusion

In this chapter we have presented formal notation, semantics, a representation scheme, and inference algorithms that facilitate the probabilistic evaluation of counterfactual queries. World knowledge is represented in the language of modified causal networks, whose root nodes are unobserved, and correspond to possible functional mechanisms operating among families of observables. The prior probabilities of these root nodes are updated by the factual information transmitted with the query, and remain fixed thereafter. The antecedent of the query is interpreted as a proposition that is established by an external intervention, thus pruning the corresponding links from the network and facilitating standard Bayesian-network computation to determine the probability of the consequent.

The algorithm has not been implemented, but, given a subjective prior distribution over the response variables, there are no new computational tasks introduced by this formalism, and the inference process follows the standard techniques for computing beliefs in Bayesian networks [Pea88]. If prior distributions over the relevant response-function variables cannot be assessed, there are methods that use the standard conditional-probability specification of Bayesian networks to compute upper and lower bounds on counterfactual probabilities. Chapter 3 will formally develop these methods.

The semantics and methodology introduced in this chapter can be adopted to nonprobabilistic formalisms as well, as long as they support two essential components: abduction (to abduce plausible functional mechanisms from the factual observations) and causal projection (to infer the consequences of the intervention-like antecedent). We should note, though, that the license to keep the response-

function variables constant stems from a unique feature of counterfactual queries, where the factual observations are presumed to occur not earlier than the counterfactual intervention. In general, when an observation takes place before an intervention, constancy of response functions would be justified if the environment remains relatively static between the observation and the intervention (e.g., if the disturbance terms ϵ_i represent unknown pre-intervention conditions). However, in a dynamic environment subject to stochastic shocks a full temporal analysis using temporally-indexed networks may be warranted or, alternatively, a canonical model of persistence should be invoked [Pea93d].

CHAPTER 3

Bounding counterfactual probabilities

3.1 Introduction

In Chapter 2, an algorithm was presented for evaluating the unique quantitative solutions to counterfactual queries when a functional model is available. However, it is rare that there is sufficient knowledge about a system's underlying mechanisms to generate a complete functional model. This chapter is concerned with the evaluation of counterfactual probabilities when this model is incomplete.

Section 3.2 describes how counterfactual probabilities may be uniquely expressed in terms of a functional model's distribution of response-functions. In Section 3.3 we will describe how these response-function distributions are constrained by a probabilistic specification over the observable variables in the system, and how the expression for the counterfactual probability may be minimized and maximized over these constraints. When the expression to be optimized is a linear function of the response-function distributions, the evaluation of bounds on the counterfactual probability may be guaranteed; however, as will be demonstrated in Section 3.4 many counterfactual probabilities are polynomial functions of the response-function distributions in which case the potential for local optima usually means that determination of bounds is not guaranteed. Finally, in Section 3.5 we demonstrate that marginalization of variables from a probabilistic causal model prior to evaluating bounds on counterfactual probabilities lead to looser bounds than if the analysis were performed on the original model.

3.2 Expressing counterfactual probabilities in terms of response-function distributions

Given the functional specification of a causal system as described in Section 2.3, we can derive an expression for a counterfactual probability $P(c^*|\hat{a}^*, o)$ in terms of the underlying functional model's parameters.

Let $\mathbf{r} = (r_{x_1}, r_{x_2}, \ldots, r_{x_n})$ represent the set of response-function variables for

the corresponding observable variables in the model. Given the value of $\mathbf{r}$, all variables $X_i \in X$ are functionally determined according to the recursive function:

$$\begin{aligned} x_i &= g_{x_i}(\mathbf{r}) \\ &= f_{x_i}(g_{u_1}(\mathbf{r}), g_{u_2}(\mathbf{r}), \ldots, g_{u_k}(\mathbf{r}), r_{x_i}) \end{aligned}$$

where $\text{pa}(X_i) = \{U_1, U_2, \ldots, U_k\} \subset X$ are the causal influences of X_i in the model.

If a set of variables $A \subset X$ in the model are externally forced to the value a, then according to the intervention-based semantics of [Pea93a], the recursive function becomes

$$\begin{aligned} x_i &= g^{\hat{a}}_{x_i}(\mathbf{r}) \\ &= \begin{cases} \hat{x}_i & \text{if } X_i \in A \\ f_{x_i}(r_{x_i}) & \text{if } X_i \notin A \text{ and } \text{pa}(X_i) = \emptyset \\ f_{x_i}(g^{\hat{a}}_{u_1}(\mathbf{r}), g^{\hat{a}}_{u_2}(\mathbf{r}), \ldots, g^{\hat{a}}_{u_k}(\mathbf{r}), r_{x_i}) & \text{otherwise} \end{cases} \end{aligned}$$

The counterfactual probability $P(c^*|\hat{a}^*, o)$ may be rewritten

$$P(c^*|\hat{a}^*, o) = \frac{P(c^*, o|\hat{a}^*)}{P(o|\hat{a}^*)}$$

Since an intervention can only affect its descendants in the graph [Pea94b] we have $P(o|\hat{a}) = P(o)$ which is readily computed from the probabilistic specification.

$P(c^*, o|\hat{a}^*)$ may be evaluated in terms of the functional model by summing the probabilities of the response-function configurations which are consistent with the arguments (c^*, $\hat{a}^*$, o). Formally,

$$P(c^*, o|\hat{a}^*) = \sum_{\mathbf{r} \in R} P(\mathbf{r})$$

where

$$R = \{\mathbf{r} | \forall_{x_i \in o}[g_{x_i}(\mathbf{r}) = x_i] \text{ and } \forall_{x^*_j \in c^*}[g^{\hat{a}}_{x_j}(\mathbf{r})] = x^*_j\}$$

Hence, the counterfactual probability may be written in terms of the structure $\{\text{pa}(x_i)\}$ and parameters $P(\mathbf{r})$ of the functional model:

$$P(c^*|\hat{a}^*, o) = \frac{\sum_{\mathbf{r} \in R} P(\mathbf{r})}{P(o)} \tag{3.1}$$

The next section will describe how the right-hand side of Eq.(3.1) may be optimized subject to the constraints imposed by the probabilistic specification. But first, we will derive an expression for the probability that Bob would have fired his rifle, if the Captain were to have given the order to shoot, given that the Captain gave the order to release the traitor and Bob did not shoot ($P(b_1^*|\hat{c}_1^*, c_0, b_0)$).

The connection between the factual and counterfactual worlds was discussed in Chapter 2 where it was argued that the response-function variables should assume the same values in both worlds. For the firing-squad example, this invariance allows the response function variables r_c and r_b to be shared between the networks corresponding to the two worlds (see Figure 3.1).

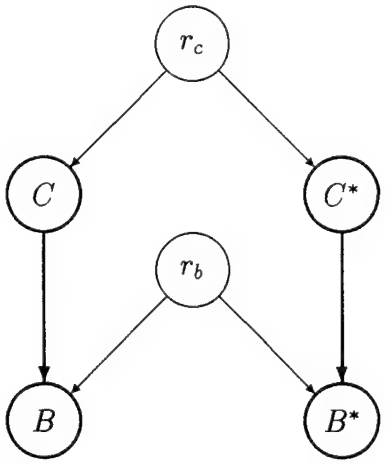

Figure 3.1: *Factual* (C, B) *and counterfactual* (C^*, B^*) *worlds for the functional analysis of the structure* $C \rightarrow B$. *The response-function variables* r_c *and* r_b *(summarizing all exogenous influences on* C *and* B*) attain the same value in the real and counterfactual worlds.*

The domain of B's response-function variable r_b is defined by Eq. (2.3), while the response-function variable for C has a two-valued domain $r_c \in \{0, 1\}$ with the following functional specification:

$$c \;=\; f_c(r_c) = h_{c,r_c}() \tag{3.2}$$

where the mappings defined by each response function $h_{c,r_c}()$ are given by

$$\begin{aligned} h_{c,0}() &= c_0 \\ h_{c,1}() &= c_1 \end{aligned}$$

One quickly notes that the prior probability distribution on r_c will be the same as the distribution over C (this is true in general for variables that were root nodes in the original causal structure):

$$\begin{aligned} P(r_c{=}0) &= P(c_0) \\ P(r_c{=}1) &= P(c_1) \end{aligned}$$

From Eq. (3.1),

$$P(b_1^*|\hat{c}_1^*, c_0, b_0) = \frac{\sum_{\mathbf{r}\in R} P(\mathbf{r})}{P(c_0, b_0)}$$

where

$$R = \{(r_c, r_b)|g_c(r_c, r_b){=}c_0 \wedge g_b(r_c, r_b){=}b_0 \wedge g_b^{\hat{c}_1}(r_c, r_b){=}b_1\} \tag{3.3}$$

The only tuple satisfying the condition in Eq. (3.3) is

$$(r_c, r_b) = (0, 1)$$

Therefore,

$$P(b_1^*|\hat{c}_1^*, c_0, b_0) = \frac{P(r_c{=}0)P(r_b{=}1)}{P(c_0, b_0)}$$

But, $P(r_c{=}0) = P(c{=}c_0)$, hence,

$$P(b_1^*|\hat{c}_1^*, c_0, b_0) = \frac{P(r_b{=}1)}{P(b_0|c_0)}$$

3.3 Constraints and optimization

The probabilistic specification $P(x_i|\mathrm{pa}(x_i))$ for a complete model imposes a set of constraints on the distribution of response functions $P(r_{x_i})$ of the form

$$P(x_i|\mathrm{pa}(x_i)) = \sum_{r_{x_i}} P(r_{x_i}) t(r_{x_i}; x_i, \mathrm{pa}(x_i)) \tag{3.4}$$

where the characteristic function t indicates which response functions r_{x_i} map the given value of X_i's causal influences $\mathrm{pa}(x_i)$ to X_i's given value x_i, i.e.

$$t(r_{x_i}; x_i, \mathrm{pa}(x_i)) = \begin{cases} 1 & \text{if } x_i = f_{x_i}(\mathrm{pa}(x_i), r_{x_i}) \\ 0 & \text{otherwise} \end{cases}$$

For an incomplete model, where X_i and X_j are assumed to have an exogenous common cause, the common constraint for these two variables will be given instead by

$$P(x_i, x_j|\mathrm{pa}_{X-\{X_j\}}(x_i), \mathrm{pa}_{X-\{X_i\}}(x_j)) = \sum_{r_{x_i}, r_{x_j}} P(r_{x_i}, r_{x_j}) t(r_{x_i}; x_i, \mathrm{pa}(x_i)) t(r_{x_j}; x_j, \mathrm{pa}(x_j)) \tag{3.5}$$

Note that the constraints in Eq. (3.5) are linear in $P(r_{x_i}, r_{x_j})$.

As an example, the constraints in the firing-squad story (which is complete with two binary variables C and B) are given by

$$P(b_1|c_0) = P(r_b{=}2) + P(r_b{=}3) \tag{3.6}$$

$$P(b_1|c_1) = P(r_b{=}1) + P(r_b{=}3) \tag{3.7}$$

$$P(c_1) = P(r_c{=}1) \tag{3.8}$$

Given the entire set of linear constraints and the objective function from Eq. (3.1), the bounds may be evaluated using techniques for optimizing non-linear objective functions under linear constraints [Sca85]. In general, the optimization procedure may converge to a local minima/maxima which would produce false bounds. If the objective is to prove that the counterfactual probability falls within a certain range, care must be taken to ensure that global optima are found.

If the objective function given by Eq. (3.1) is linear, the minimum/maximum may be determined using linear programming techniques. In this case, when the problem size is small enough, we may also derive closed-form bounds to the counterfactual probability in terms of the probabilistic specification. This is accomplished by enumerating the vertices in the dual linear programming problem (see Appendix B).

For the firing-squad example, the symbolic expression for the counterfactual probability $P(b_1^*|\hat{c}_1^*, c_0, b_0)$ may be optimized over the space of linear constraints given by Eqs. (3.6)–(3.8). The resulting symbolic bounds are:

$$\frac{1}{P(b_0|c_0)} \max \left\{ \begin{array}{c} P(b_1|c_1) - P(b_1|c_0) \\ 0 \end{array} \right\} \leq P(b_1^*|\hat{c}_1^*, c_0, b_0) \leq \frac{1}{P(b_0|c_0)} \min \left\{ \begin{array}{c} P(b_0|c_0) \\ P(b_1|c_1) \end{array} \right\}$$

By substituting the known conditional probabilities from Section 2.3

$$\begin{aligned} P(b_0|c_0) &= 0.90 \\ P(b_1|c_1) &= 0.90 \end{aligned}$$

we can evaluate the numeric bounds on the counterfactual probability:

$$8/9 \leq P(b_1^*|\hat{c}_1^*, c_0, b_0) \leq 1$$

Sometimes, one may feel confident in claiming that additional constraints should be imposed on the parameters defining the distribution of response-functions. For example, we may subjectively believe that Bob never confuses the shoot and release signals, which is simply written $P(r_b=2) = 0$ and added to the existing set of constraints. The optimization of the expression for the counterfactual probability then proceeds as before. In this case, this assumption is sufficient to uniquely determine the counterfactual probability

$$P(b_1^*|\hat{c}_1^*, c_0, b_0) = 8/9$$

This shows that partial knowledge or belief about the distribution of response-functions is an important technique for tightening bounds on counterfactual probabilities given only a probabilistic specification of observable variables.

3.4 Nonlinear expressions

Unfortunately, a closed-form expression for the counterfactual probability is not always a linear function of the parameters of the response-function distributions. This will be demonstrated by the following example which relates to the firing-squad example previously discussed.

First, the original model will be expanded by incorporating the additional knowledge that there is only one other rifleman, Dave, whose tendency to fire is independent of Bob's firing given the Captain's signal. The causal structure for this model is depicted in Figure 3.2. The story relating Dave's firing habits will be similar to Bob's habits described in Section 1.2. D is a deterministic function of C, and D's response-function variable r_d

$$d = f_d(c, r_d) = h_{d,r_d}(c) \tag{3.9}$$

where

$$h_{d,0}(c) = d_0 \tag{3.10}$$

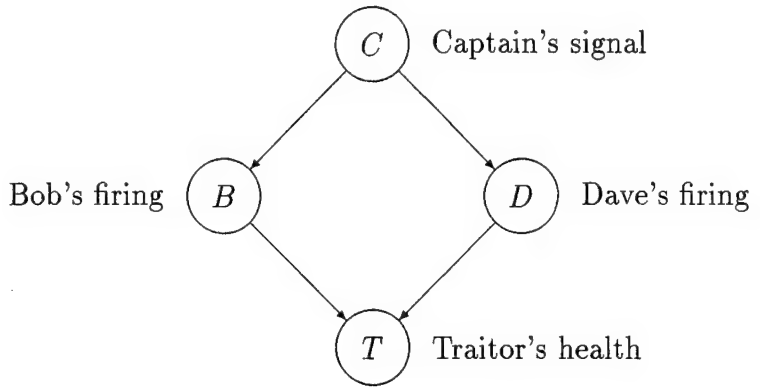

Figure 3.2: *Causal structure reflecting the influence that the Captain's signal has on Bob and Dave's firing, and the influence that their firing has on the Traitor's health.*

$$h_{d,1}(c) = \begin{cases} d_0 & \text{if } c = c_0 \\ d_1 & \text{if } c = c_1 \end{cases} \tag{3.11}$$

$$h_{d,2}(c) = \begin{cases} d_1 & \text{if } c = c_0 \\ d_0 & \text{if } c = c_1 \end{cases} \tag{3.12}$$

$$h_{d,3}(c) = d_1 \tag{3.13}$$

In addition, T is now a deterministic function of B, D, and T's response-function variable r_t

$$t = f_t(b, d, r_t) = h_{t,r_t}(b, d) \tag{3.14}$$

where

$$h_{t,0}(b,d) = t_0$$

$$h_{t,1}(b,d) = \begin{cases} t_0 & \text{if } (b,d) \neq (b_1,d_1) \\ t_1 & \text{if } (b,d) = (b_1,d_1) \end{cases}$$

$$h_{t,2}(b,d) = \begin{cases} t_0 & \text{if } (b,d) \neq (b_0,d_1) \\ t_1 & \text{if } (b,d) = (b_0,d_1) \end{cases}$$

$$h_{t,3}(b,d) = \begin{cases} t_0 & \text{if } d = d_0 \\ t_1 & \text{if } d = d_1 \end{cases}$$

$$h_{t,4}(b,d) = \begin{cases} t_0 & \text{if } (b,d) \neq (b_1,d_0) \\ t_1 & \text{if } (b,d) = (b_1,d_0) \end{cases}$$

$$h_{t,5}(b,d) = \begin{cases} t_0 & \text{if } b = b_0 \\ t_1 & \text{if } b = b_1 \end{cases}$$

$$
\begin{aligned}
h_{t,6}(b,d) &= \begin{cases} t_0 & \text{if } (b,d) \in \{(b_0,d_0),(b_1,d_1)\} \\ t_1 & \text{if } (b,d) \in \{(b_1,d_0),(b_0,d_1)\} \end{cases} \\
h_{t,7}(b,d) &= \begin{cases} t_0 & \text{if } (b,d) = (b_0,d_0) \\ t_1 & \text{if } (b,d) \neq (b_0,d_0) \end{cases} \\
h_{t,8}(b,d) &= \begin{cases} t_0 & \text{if } (b,d) \neq (b_0,d_0) \\ t_1 & \text{if } (b,d) = (b_0,d_0) \end{cases} \\
h_{t,9}(b,d) &= \begin{cases} t_0 & \text{if } (b,d) \in \{(b_1,d_0),(b_0,d_1)\} \\ t_1 & \text{if } (b,d) \in \{(b_0,d_0),(b_1,d_1)\} \end{cases} \\
h_{t,10}(b,d) &= \begin{cases} t_0 & \text{if } b = b_1 \\ t_1 & \text{if } b = b_0 \end{cases} \\
h_{t,11}(b,d) &= \begin{cases} t_0 & \text{if } (b,d) = (b_1,d_0) \\ t_1 & \text{if } (b,d) \neq (b_1,d_0) \end{cases} \\
h_{t,12}(b,d) &= \begin{cases} t_0 & \text{if } c = d_1 \\ t_1 & \text{if } c = d_0 \end{cases} \\
h_{t,13}(b,d) &= \begin{cases} t_0 & \text{if } (b,d) = (b_0,d_1) \\ t_1 & \text{if } (b,d) \neq (b_0,d_1) \end{cases} \\
h_{t,14}(b,d) &= \begin{cases} t_0 & \text{if } (b,d) = (b_1,d_1) \\ t_1 & \text{if } (b,d) \neq (b_1,d_1) \end{cases} \\
h_{t,15}(b,d) &= t_1
\end{aligned}
$$

Suppose, that we observe the Captain give the signal to shoot (c_1), Bob fires his rifle (b_1), and the Traitor survives (t_1). If the Captain had not given the signal to shoot, what is the probability that the Traitor would have died (t_0), i.e., what is $P(t_0^*|\hat{c}_0^*, c_1, b_1, t_1)$?

The instantiated graphical structure for evaluating this conditional probability is shown in Figure 3.3.

According to the procedure described in Section 3.2, we can write the probability of the counterfactual consequent in terms of the response-function distributions $P(r_c)$, $P(r_b)$, $P(r_d)$, $P(r_t)$:

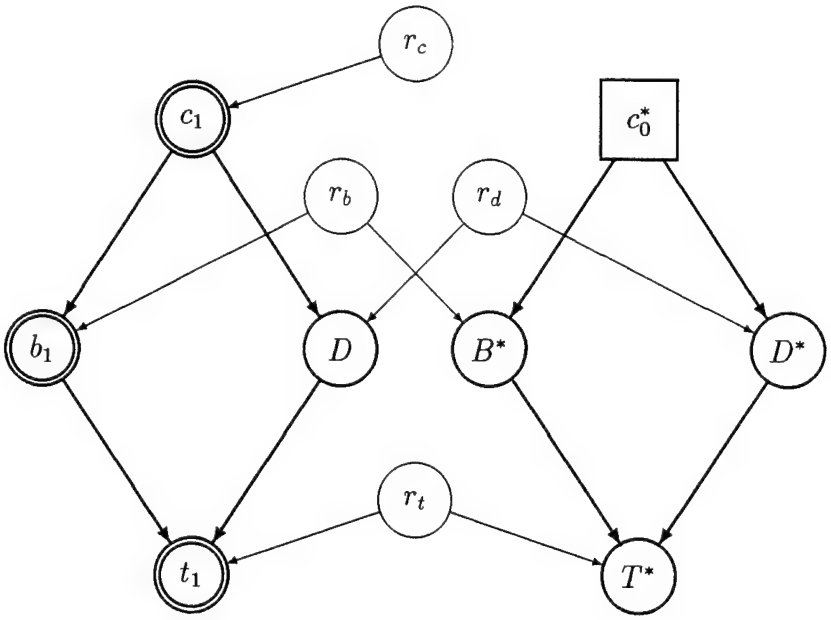

Figure 3.3: *To evaluate the counterfactual probability* $P(t_0^*|\hat{c}_0^*, c_1, b_1, t_1)$*, the combined functional model (factual/counterfactual worlds) is instantiated with observations* c_1, b_1, t_1 *and intervention* $\hat{c}_0^*$ *(links pointing to* c_0^* *are severed).*

$$P(t_0^*|\hat{c}_0^*, c_1, b_1, t_1) = \tag{3.15}$$

$$\frac{1}{P(b_1, t_1|c_1)} \left[\begin{array}{l} P(r_b{=}1) \left[\begin{array}{l} P(r_d{=}0) \sum_{k\in\{4,5,6,7\}} P(r_t{=}k) + \\ P(r_d{=}1) \sum_{k\in\{1,3,5,7\}} P(r_t{=}k) + \\ P(r_d{=}2) \sum_{k\in\{4,5,12,13\}} P(r_t{=}k) + \\ P(r_d{=}3) \sum_{k\in\{1,5,9,13\}} P(r_t{=}k) \end{array} \right] + \\ P(r_b{=}3) \left[\begin{array}{l} P(r_d{=}1) \sum_{k\in\{1,3,9,11\}} P(r_t{=}k) + \\ P(r_d{=}2) \sum_{k\in\{4,6,12,14\}} P(r_t{=}k) \end{array} \right] \end{array} \right]$$

with the following constraints over the response-functions' distributions:

$$\begin{aligned} P(r_b{=}0) + P(r_b{=}1) &= P(b_0|c_0) \\ P(r_b{=}2) + P(r_b{=}3) &= P(b_1|c_0) \end{aligned}$$

$$\begin{aligned} P(r_b{=}0) + P(r_b{=}2) &= P(b_0|c_1) \\ P(r_b{=}1) + P(r_b{=}3) &= P(b_1|c_1) \end{aligned}$$

$$
\begin{aligned}
P(r_d{=}0) + P(r_d{=}1) &= P(d_0|c_0) \\
P(r_d{=}2) + P(r_d{=}3) &= P(d_1|c_0)
\end{aligned}
$$

$$
\begin{aligned}
P(r_d{=}0) + P(r_d{=}2) &= P(d_0|c_1) \\
P(r_d{=}1) + P(r_d{=}3) &= P(d_1|c_1)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i\in\{0,1,2,3,4,5,6,7\}} P(r_t{=}i) &= P(t_0|b_0, d_0) \\
\sum_{i\in\{8,9,10,11,12,13,14,15\}} P(r_t{=}i) &= P(t_1|b_0, d_0)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i\in\{0,1,4,5,8,9,12,13\}} P(r_t{=}i) &= P(t_0|b_0, d_1) \\
\sum_{i\in\{2,3,6,7,10,11,14,15\}} P(r_t{=}i) &= P(t_1|b_0, d_1)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i\in\{0,1,2,3,8,9,10,11\}} P(r_t{=}i) &= P(t_0|b_1, d_0) \\
\sum_{i\in\{4,5,6,7,12,13,14,15\}} P(r_t{=}i) &= P(t_1|b_1, d_0)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i\in\{0,2,4,6,8,10,12,14\}} P(r_t{=}i) &= P(t_0|b_1, d_1) \\
\sum_{i\in\{1,3,5,7,9,11,13,15\}} P(r_t{=}i) &= P(t_1|b_1, d_1)
\end{aligned}
$$

Eq. (3.15) is a polynomial expression of the response-function variables, and is therefore not directly amenable to the linear-optimization procedure detailed in Appendix B. As an alternative, one could apply techniques for optimizing nonlinear functions in a convex polytope [Sca85]; however, there is no guarantee that the global optima will be found by these procedures, so care must be taken in interpreting the results. In Chapter 4, we will find that global optima are guaranteed when computing counterfactual beliefs in an order-of-magnitude probability calculus.

There are some cases, however, where a polynomial expression is amenable to linear optimization, because the expression may be manipulated into a form where linear sub-expressions may be optimized independently. Once these sub-expressions are optimized, then their optimal values may be substituted into the original expression, and the procedure is repeated until we are left with a linear

expression that is directly optimizable.

Theorem 3.4.1 *Several linear expressions* $f_1(\vec{x})$, $f_2(\vec{x})$, ..., $f_n(\vec{x})$ *may be* independently optimized *if*

$$\sum_k opt f_k(\vec{x}) \quad = \quad opt[\sum_k f_k(\vec{x})]$$

For example, suppose that we have a model for our domain containing three binary variables A, B, and C, with the structure $A \rightarrow B \rightarrow C$ and the conditional probability distributions $P(a)$, $P(b|a)$, and $P(c|b)$. We make the observation $\{a_0, c_0\}$ and then wish to know $P(c_1^*|\hat{a}_1^*, a_0, c_0)$, i.e., the probability that C would have been c_1, if A were a_1. Figure 3.4 shows the structure of the functional model corresponding to the probabilistic specification.

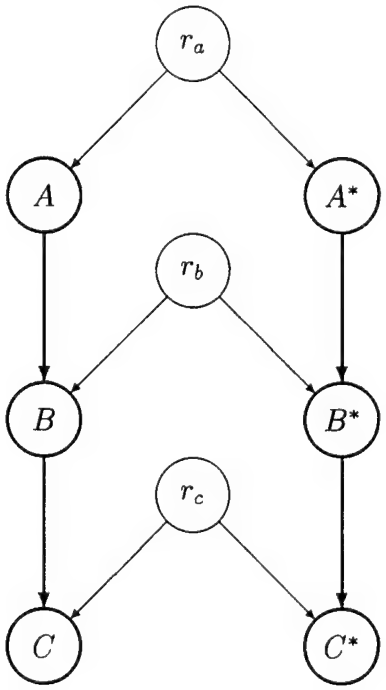

Figure 3.4: *Bayesian model for evaluating counterfactual queries when the causal structure is given by* $A \rightarrow B \rightarrow C$.

C is a deterministic function of B and r_c, and B is a deterministic function of A and r_b in the complete model. In terms of this models' response-function distributions, we may express the counterfactual probability:

$$P(c_1^*|\hat{a}_1^*, a_0, c_0) \quad = \quad \frac{P(r_b{=}1)P(r_c{=}1) + P(r_b{=}2)P(r_c{=}2)}{P(c_0|a_0)} \qquad (3.16)$$

This expression is nonlinear with respect to the response-function distributions $P(r_b)$ and $P(r_c)$; however, applying Theorem 3.4.1 shows that $P(r_b=1)$ and $P(r_b=2)$ may be optimized independently.

$$\max\left\{\begin{array}{c} 0 \\ P(b_1|a_1) - P(b_1|a_0) \end{array}\right\} \leq P(r_b{=}1) \leq \max\left\{\begin{array}{c} P(b_0|a_0) \\ P(b_1|a_1) \end{array}\right\} \quad (3.17)$$

$$\max\left\{\begin{array}{c} 0 \\ P(b_1|a_0) - P(b_1|a_1) \end{array}\right\} \leq P(r_b{=}2) \leq \max\left\{\begin{array}{c} P(b_0|a_1) \\ P(b_1|a_0) \end{array}\right\} \quad (3.18)$$

$$\begin{aligned} &\max\left\{\begin{array}{c} P(b_1|a_1) - P(b_1|a_0) \\ P(b_1|a_0) - P(b_1|a_1) \end{array}\right\} \\ &\qquad \leq P(r_b{=}1) + P(r_b{=}2) \leq \\ &\qquad \max\left\{\begin{array}{c} P(b_0|a_0) + P(b_0|a_1) \\ P(b_1|a_0) + P(b_1|a_1) \end{array}\right\} \end{aligned} \quad (3.19)$$

Summing the right hand side of Eqs. (3.17) and (3.18):

$$\begin{aligned} &\max\left\{\begin{array}{c} P(b_0|a_0) \\ P(b_1|a_1) \end{array}\right\} + \max\left\{\begin{array}{c} P(b_0|a_1) \\ P(b_1|a_0) \end{array}\right\} = \\ &\qquad \max\left\{\begin{array}{c} P(b_0|a_0) + P(b_0|a_1) \\ P(b_1|a_0) + P(b_1|a_1) \\ P(b_0|a_0) + P(b_1|a_0) = 1 \\ P(b_0|a_1) + P(b_1|a_1) = 1 \end{array}\right\} \end{aligned} \quad (3.20)$$

But one of the first two terms in the right hand side of Eq. (3.20) must be greater than or equal to one, while the other term is less than or equal to one; therefore, the expression reduces to the right hand side of Eq. (3.19). Similar arguments lead to the conclusion that the sum of the left hand sides in Eqs. (3.17) and (3.18) is equal to the left hand side of Eq. (3.19). The conditions in Theorem 3.4.1 are satisfied allowing us to use linear optimization to compute the bounds on the counterfactual probability:

$$\begin{aligned} &\frac{1}{P(c_0|a_0)} \max\left\{\begin{array}{c} 0 \\ (P(b_1|a_1) - P(b_1|a_0))(P(c_1|b_1) - P(c_1|b_0)) \\ (P(b_1|a_0) - P(b_1|a_1))(P(c_1|a_0) - P(c_1|a_1)) \end{array}\right\} \\ &\qquad \leq P(c_1^*|\hat{a}_1^*, a_0, c_0) \leq \\ &\qquad \frac{1}{P(c_0|a_0)} \min\left\{\begin{array}{c} P(b_0|a_0)P(c_0|b_0) + P(b_0|a_1)P(c_0|b_1) \\ P(b_1|a_1)P(c_0|b_0) + P(b_1|a_0)P(c_0|b_1) \\ P(b_0|a_0)P(c_1|b_1) + P(b_0|a_1)P(c_1|b_0) \\ P(b_1|a_1)P(c_1|b_1) + P(b_1|a_0)P(c_1|b_1) \end{array}\right\} \end{aligned} \quad (3.21)$$

This section has shown that although many interesting counterfactual probabilities are polynomial with respect to the underlying response-function distribution, and hence susceptible to the problem arising from local optima within the parameter space, there are some cases where linear optimization is still possible because some of the terms in the expression may be independently optimized, and then combined to form a closed-form expression for the bounds. This technique will be applied in Chapter 5 when deriving bounds on treatment effects given a subject's category of treatment consumption.

3.5 Model marginalization

In the last section, we computed the bounds for a counterfactual probability based on a model containing three variables A, B, and C, where B's value did not take part in the specification of the counterfactual query. One might consider marginalizing B out of the model, because B is not referenced in our observations or the counterfactual conditional. Although this is admissible when the prior probabilities on the response-function variables $P(r_b)$ and $P(r_c)$ are known (allowing exact calculation of the counterfactual probability), this strategy is fallible when these distributions are unspecified and hence only bounds on the the counterfactual probability may be computed.

Figure 3.4 shows the structure of the functional model corresponding to the probabilistic specification. If we marginalize out the variable B, the structure of the functional model is given by Figure 3.5 (note that the response-function variable for C in the partial model is denoted by s_c). The mapping from the complete model's conditional probability specifications to the partial model's specification is given simply by

$$P(c_1|a_0) = P(c_1|b_0)P(b_0|a_0) + P(c_1|b_1)P(b_1|a_0) \quad (3.22)$$

$$P(c_1|a_1) = P(c_1|b_0)P(b_0|a_1) + P(c_1|b_1)P(b_1|a_1) \quad (3.23)$$

In the complete model, C is a deterministic function of B and r_c, and B is a deterministic function of A and r_b. However, in the partial model, C is a deterministic function of A and s_c. In terms of the partial models' response-function distributions, we may express the counterfactual probability:

$$P(c_1^*|\hat{a}_1^*, a_0, c_0) = \frac{P(s_c{=}1)}{P(c_0|a_0)} \quad (3.24)$$

Given an instantiation of B and C's response-function distributions, the two

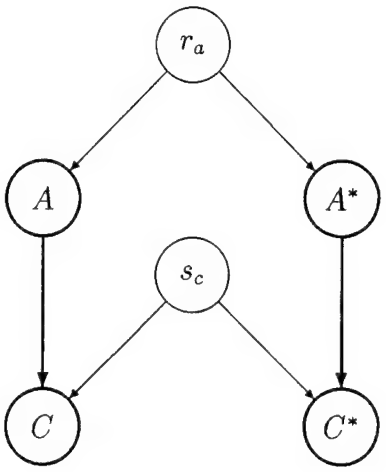

Figure 3.5: *Partial model over A and C.*

counterfactual probabilities given by Eqs. (3.16) and (3.24) will always be equal, because the numerator of Eq. (3.24) is just the result of marginalizing out B.

However, in terms of the partial model, the bounds on the counterfactual probability are derived as:

$$\frac{1}{P(c_0|a_0)} \max \left\{ \begin{array}{c} 0 \\ P(c_1|a_1) - P(c_1|a_0) \end{array} \right\} \leq P(c_1^*|\hat{a}_1^*, a_0, c_0) \leq \frac{1}{P(c_0|a_0)} \min \left\{ \begin{array}{c} P(c_0|a_0) \\ P(c_1|a_1) \end{array} \right\}$$

which may be expanded as follows using Eqs. (3.22) and (3.23)

$$\frac{1}{P(c_0|a_0)} \max \left\{ \begin{array}{c} 0 \\ [P(b_0|a_1)P(c_1|b_0) + P(b_1|a_1)P(c_1|b_1) - \\ \qquad P(b_0|a_0)P(c_0|b_0) + P(b_1|a_0)P(c_0|b_1)] \end{array} \right\} \leq P(c_1^*|\hat{a}_1^*, a_0, c_0) \leq \frac{1}{P(c_0|a_0)} \min \left\{ \begin{array}{c} P(b_0|a_0)P(c_0|b_0) + P(b_1|a_0)P(c_0|b_1) \\ P(b_0|a_1)P(c_1|b_0) + P(b_1|a_1)P(c_1|b_1) \end{array} \right\}$$

Comparing these bounds to those computed with the full model, Eq. (3.21), one can see that the numeric bounds evaluated from the partial model are never tighter and almost always looser than those evaluated from the complete model analysis.

In this section, we have demonstrated that unreferenced variables may not be marginalized out of the probabilistic causal model without potentially affecting the evaluation of bounds for a counterfactual probability. However, the bounds evaluated from the marginalized model still hold true — they are just not necessarily tight. Therefore, one may still consider evaluating bounds under the marginalized model if one cannot guarantee that global optima have been found from the analysis using the complete model.

3.6 Conclusion

This chapter has developed a procedure for evaluating bounds on counterfactual probabilities. The corner-stone of counterfactual analysis is the use of functional models with response-function variables, for which the counterfactual probability may be uniquely written. The task of determining bounds involves the optimization of this expression under the constraints imposed by the known probabilistic specification. In general, the task is reduced to the optimization of a polynomial function subject to linear constraints, which introduces the problem of local minima/maxima. However, if the counterfactual probability is linear with respect to the functional specification, then the bounds are easily found via linear programming. In addition, in some cases we may be able to derive closed-form bounds on counterfactual probabilities in terms of the probabilistic specification.

CHAPTER 4

Evaluating counterfactuals from κ rankings: Computation and Bounds

4.1 Introduction

In Chapters 2 and 3 a formalism was developed for evaluating and bounding counterfactual probabilities given a causal structure of the relevant domain along with conditional probabilities of each variable given its set of causal influences. Detractors of reasoning with probabilistic causal networks claim that it is unreasonable to assume that we can obtain the numbers which parameterize the causal model, and that we may only elicit crude measures of belief from human reasoners. An alternative representation of these belief measures is given by an order-of-magnitude abstraction of probabilities, known as κ-rankings [Spo88].

In general, the objective function to optimize for evaluating bounds on counterfactual probabilities will be a polynomial function with respect to the unspecified prior probabilities on the response-function variables. Therefore, algorithms for optimizing cannot always verify that the global minima/maxima has been discovered, because the algorithms may terminate at local minima/maxima. If we cannot guarantee global optima, then the returned bounds on the counterfactual probability are too tight; and therefore, are not bounds. If, however, we represent knowledge by a κ ranking over the worlds, then we can always evaluate the upper and lower bounds on our belief in a counterfactual consequent. Of course, this only gives us an approximation to the bounds that would be determined using a fully specified probabilistic causal model.

In this chapter, we will reformulate the evaluation of bounds on counterfactual beliefs in terms of κ rankings over possible worlds. The next section will give background on reasoning with κ ranking functions. In Section 4.4, a general description of a procedure for evaluating bounds on counterfactual beliefs given κ ranking functions will be given. Section 4.5 will demonstrate this on an example, and finally Section 4.6 will give some concluding remarks.

4.2 κ rankings

κ rankings ([Spo88]) provide an order-of-magnitude abstraction of probability distributions that states that if $P(a)$ is of order $O(e^n)$ for some constant e less than 1 and non-negative integer n, than the κ ranking of a is $\kappa(a) = n$. Note that as a probability decreases, its κ ranking will increase. This transformation from probabilities to κ rankings partitions the range of probabilities into equivalence classes designated by a non-negative integer that indicates how surprising a particular event would be ($\kappa(a) = 0$ indicates that event a would not be surprising).

One of the obvious benefits of κ rankings is greater ease in specifying beliefs: rather than specifying a precise probability, only a crude estimate of the probability is necessary. Of course, this also means that the accuracy of the result is less precise; if you do not have precise probabilities to begin with, the solution should not be expected to be precise.

The basic operators of ranking functions correspond very nicely with the operators in probability theory: multiplication and addition in probability theory are replaced by addition and minimization, respectively, in κ calculus. While probability theory has the following axioms

$$P(a) = \sum_{w \models a} P(w) \tag{4.1}$$

$$P(a) + P(\neg a) = 1 \tag{4.2}$$

$$P(a, b) = P(b|a)P(a) \tag{4.3}$$

κ calculus has the corresponding set of axioms

$$\kappa(a) = \min_{w \models a} \kappa(w) \tag{4.4}$$

$$\min\{\kappa(a), \kappa(\neg a)\} = 0 \tag{4.5}$$

$$\kappa(a, b) = \kappa(b|a) + \kappa(a) \tag{4.6}$$

We will now use this relationship between κ and probability calculi to describe the procedure for evaluate counterfactual κ rankings.

4.3 κ ranked counterfactuals

Consider the causal structure $C \rightarrow B$ with an associated conditional kappa ranking $\kappa(b|c)$. Suppose that we have observed (c_0, b_0). What is our belief that B

would have been equal to b_1, if C were c_1. According to the formalism for evaluating counterfactual probabilities in Chapter 2, we generate a functional model for the given causal structure. In this case we introduce a response-function variable r_b which specifies the mapping from C to B as follows:

$$b = f_b(c, r_b) = h_{b,r_b}(c)$$

where

$$\begin{aligned} h_{b,0}(c) &= b_0 \\ h_{b,1}(c) &= \begin{cases} b_0 & \text{if } c = c_0 \\ b_1 & \text{if } c = c_1 \end{cases} \\ h_{b,2}(c) &= \begin{cases} b_1 & \text{if } c = c_0 \\ b_0 & \text{if } c = c_1 \end{cases} \\ h_{b,3}(c) &= b_1 \end{aligned}$$

The kappa ranking over these response functions $\kappa(r_b)$ then parameterizes the model.

Similar to the strategy developed in 3 we can write the *counterfactual kappa ranking* for our query as follows:

$$\begin{aligned} \kappa(b_1^*|\hat{c}_1^*, c_0, b_0) &= \kappa(c_0, b_0, b_1^*|\hat{c}_1^*) - \kappa(c_0, b_0) \\ &= \kappa(r_{b1}) - \kappa(b_0|c_0) \end{aligned}$$

If the kappa ranking over the response-function variable r_b is known, then a unique counterfactual kappa rank may be computed; however, if this information is not available, then the counterfactual kappa ranking may only be bounded under the constraints given by the known kappa ranking $\kappa(b|c)$. These constraints are:

$$\kappa(b_0|c_1) = \min\{\kappa(r_{b0}), \kappa(r_{b2})\} \tag{4.7}$$

$$\kappa(b_0|c_0) = \min\{\kappa(r_{b0}), \kappa(r_{b1})\} \tag{4.8}$$

$$\kappa(b_1|c_0) = \min\{\kappa(r_{b2}), \kappa(r_{b3})\} \tag{4.9}$$

$$\kappa(b_1|c_1) = \min\{\kappa(r_{b1}), \kappa(r_{b3})\} \tag{4.10}$$

$$\kappa(r_{bj}) \geq 0 \qquad \forall j \in \{0, 1, 2, 3\} \tag{4.11}$$

The formulation of the problem was straight-forward following the formalism of Chapter 2; however, an appropriate mechanism needs to be available for

performing this integer optimization with constraints containing minimization operators.

Eqs. (4.7)–(4.10) immediately imply

$$\kappa(r_{b0}) \geq \max\{\kappa(b_0|c_0); \kappa(b_0|c_1)\} \tag{4.12}$$

$$\kappa(r_{b1}) \geq \max\{\kappa(b_0|c_0); \kappa(b_1|c_1)\} \tag{4.13}$$

$$\kappa(r_{b2}) \geq \max\{\kappa(b_1|c_0); \kappa(b_0|c_1)\} \tag{4.14}$$

$$\kappa(r_{b3}) \geq \max\{\kappa(b_1|c_0); \kappa(b_1|c_1)\} \tag{4.15}$$

No dependencies exist among these expressions that prevent equality from holding in all these constraints; therefore, finding the minimum for individual $\kappa(r_b)$ terms is trivial. For our counterfactual query, this leads to a lower bound on the counterfactual κ-ranking:

$$\kappa(b_1^*|\hat{c}_1^*, c_0, b_0) \geq \max\left\{ \begin{array}{c} 0 \\ \kappa(b_1|c_1) - \kappa(b_0|c_0) \end{array} \right\} \tag{4.16}$$

When maximizing $\kappa(r_{b1})$, there are only two situations to consider: either $\kappa(r_{b1})$ is completely unbounded from above; or the upper bound is equal to the lower bound in Eq. (4.13), i.e.,

$$\kappa(r_{b1}) = \max\{\kappa(b_0|c_0); \kappa(b_1|c_1)\} \tag{4.17}$$

To determine which situation holds, we first remove $\kappa(r_{b1})$ from the minimization sets of Eqs. (4.7)–(4.10), and check for satisfiability. If satisfied, then $\kappa(r_{b1})$ is not bounded from above; otherwise, the bounds reduce to equality

$$\kappa(b_1^*|\hat{c}_1^*, c_0, b_0) = \max\left\{ \begin{array}{c} 0 \\ \kappa(b_1|c_1) - \kappa(b_0|c_0) \end{array} \right\}$$

Of course, if the counterfactual κ is equal to zero, then we would like to know what the counterfactual κ is for the negation of the counterfactual consequent. For our example, we would be interested in $\kappa(b_0^*|\hat{c}_1^*, c_0, b_0)$. Applying the same procedure as before we obtain the lower bound

$$\kappa(b_0^*|\hat{c}_1^*, c_0, b_0) \geq \max\left\{ \begin{array}{c} 0 \\ \kappa(b_0|c_1) - \kappa(b_0|c_0) \end{array} \right\}$$

If $\kappa(r_{b0}) \longrightarrow \infty$ is not satisfiable, then the bounds reduce to equality

$$\kappa(b_0^*|\hat{c}_1^*, c_0, b_0) = \max\left\{ \begin{array}{c} 0 \\ \kappa(b_0|c_1) - \kappa(b_0|c_0) \end{array} \right\}$$

4.4 General case

4.4.1 Functional expression

In Section 3.2, we gave a declarative definition for counterfactual probabilities written in terms of the structure $\{\text{pa}(x_i)\}$ and the parameters of the response-function distributions:

$$P(c^*|\hat{a}^*, o) = \frac{\sum_{\mathbf{r}\in R} P(\mathbf{r})}{P(o)}$$

where

$$R = \{\mathbf{r}|\forall_{x_i\in o}[x_i = f_{x_i}(\mathbf{r})] \text{ and } \forall_{x_j^*\in c^*}[x_j^* = g_{x_j}^{\hat{a}}(\mathbf{r})]\}$$

This definition may be transformed according to Eqs. (4.1)–(4.6) into an expression written in terms of the κ ranking functions of the response-function variables:

$$\kappa(c^*|\hat{a}^*, o) = \min_{\mathbf{r}\in R} \kappa(\mathbf{r}) - \kappa(o) \tag{4.18}$$

where the form of $\kappa(\mathbf{r})$ is always given by the sum of κ's for each independent set of response function variables in $\mathbf{r}$.

4.4.2 Constraints

The κ rankings over the model's observable variables $\kappa(x_i|\text{pa}(x_i))$ impose a set of constraints on the κ ranking over the response-function variables $\kappa(r_{x_i})$ of the form

$$\kappa(x_i|\text{pa}(x_i)) = \min\{\kappa(r_{x_i}) : x_i = f_{x_i}(\text{pa}(x_i), r_{x_i})\} \tag{4.19}$$

Similar to the treatment of exogenous common causes discussed in Section 3.3, if X_i and X_j are assumed to have an exogenous common cause, then the common constraint for these two variables will be given instead by

$$\begin{aligned}&\kappa(x_i, x_j|\text{pa}(x_i) - \{x_j\}, \text{pa}(x_j) - \{x_i\}) \\ &= \min\{\kappa(r_{x_i,x_j}) : x_i = f_{x_i}(\text{pa}(x_i), r_{x_i}) \text{ and } x_j = f_{x_j}(\text{pa}(x_j), r_{x_j})\}\end{aligned} \tag{4.20}$$

Therefore, in general, we will be optimizing a function with minimization operators over constraints also containing minimization operators.

4.4.3 Optimization

In this section we will show that optimization of the objective κ function (Eq. (4.18) under the constraints of Eqs. (4.19) and (4.20) is trivial for the minimum value, but requires either a complete enumeration or a search procedure for determining the upper bound on the κ-ranking.

4.4.3.1 Minimization

The constraints given by Eq. (4.19) immediately imply the following lower bound on the κ ranking of individual response functions:

$$\kappa(r_{x_i}{=}j) \;\geq\; \max\{\kappa(x_i|\mathrm{pa}(x_i)) : x_i = f_{x_i}(\mathrm{pa}(x_i), r_{x_i}{=}j)\} \tag{4.21}$$

These lower bounds are obtained simply by substituting the known conditional κ rankings $\kappa(x_i|\mathrm{pa}(x_i))$ into the right hand side of Eq. (4.21). Given that the objective function consists only of minimization and summation operators, the strict lower bound on the κ of the counterfactual can always be evaluated by substituting in the lower bounds for each $\kappa(r_{x_i})$.

4.4.3.2 Maximization

In maximizing the objective function, it helps to note that if a response-function rank $\kappa(r_{x_i}{=}j)$ is not forced to be equal to its lower bound given by Eq. (4.21), then that κ term is completely unbounded from above, and may be assumed to be infinite. Therefore, when we try to maximize the objective function, we will set each response-function κ to either its lower bound value or infinite (which is equivalent to removing every instance of that κ term from the objective function.

This suggests a crude algorithm for evaluating the upper bound on the objective function: simply evaluate the objective function for every configuration of response-function κ's consistent with the constraints imposed by the known conditional κ rankings, and take the maximum. Besides the enumeration of every configuration, we must also have a means for checking the consistency of each configuration. Although computationally expensive, if all configurations can be enumerated, we can guarantee that the strict upper bound may be computed.

There are ways in which we may decrease the computational cost by performing some preprocessing to eliminate a majority of configurations from the search space. In addition, the search space can be sorted to speed up the task. A

formal algorithm will not be presented, but an example demonstrating the main concepts of maximizing a counterfactual κ will be discussed in the next section.

4.5 Example

In Section 3.4, we attempted to evaluate a counterfactual query related to the firing-squad example. In order to evaluate bounds on the counterfactual probability, a polynomial objective function over a set of linear constraints was to be optimized. Unfortunately, methods for optimizing polynomial functions are plagued by the presence of local minima/maxima in the parameter space. However, if the belief specification is given in terms of κ rankings, the bounds on the counterfactual κ ranking can be determined precisely.

In order to make the κ bounds on the counterfactual conditional more interesting, the story behind the causal structure of Figure 3.3 will be changed. Suppose there are four individuals, Carol, Bob, Dave, and Tina, with a known pattern of party attendance. The variables C, B, D, and T indicate whether each individual attended the party, respectively; the values c_1, b_1, d_1, and t_1 indicating that the individuals were at the party, while c_0, b_0, d_0, and t_0 indicating that the individuals were not at the party.

Bob really dislikes parties so almost never attends them, but if Carol is there he is slightly more likely to be there than if Carol is not at the party. This can be modelled in κ rankings as follows:

$$\kappa(b_0|c_0) = 0 \qquad \kappa(b_0|c_1) = 0$$
$$\kappa(b_1|c_0) = 2 \qquad \kappa(b_1|c_1) = 1$$

Dave, though, loves parties so he almost always attends them. However, if Carol is there he is a little less likely to be there than if Carol is not there. This can be modelled as follows:

$$\kappa(d_0|c_0) = 3 \qquad \kappa(d_0|c_1) = 1$$
$$\kappa(d_1|c_0) = 0 \qquad \kappa(d_1|c_1) = 0$$

Tina is a friend of Bob and Dave and is not very excited about going to parties. She also knows that Bob and Dave get into scuffles when they get together; therefore, Tina typically will not go to parties if both Bob and Dave are going to be there. The κ ranking representing this information is given by

$$\kappa(t_0|b_0, d_0) = 0 \qquad \kappa(t_0|b_1, d_0) = 4$$

$$\kappa(t_1|b_0,d_0)=2 \qquad \kappa(t_1|b_1,d_0)=0$$

$$\kappa(t_0|b_0,d_1)=3 \qquad \kappa(t_0|b_1,d_1)=0$$
$$\kappa(t_1|b_0,d_1)=0 \qquad \kappa(t_1|b_1,d_1)=1$$

Each variable is a deterministic function of its observable causal influences and its response-function variable according to Eqs. (3.2), (2.3), (3.9), and (3.14).

Suppose that we observe Carol at the party (c_1), Bob at the party (b_1), and Tina at the party (t_1). If Carol were not at the party (c_0), how surprised would we be to see Tina absent from the party (t_0)? In other words, what is $\kappa(t_0^*|\hat{c}_0^*, c_1, b_1, t_1)$?

The instantiated graphical structure for evaluating this κ ranking is the same as that depicted in Figure 3.3.

According to the procedure described in Section 4.4.1, we can write the κ rank of the counterfactual consequent in terms of the response-function κ-rankings $\kappa(r_c), \kappa(r_b), \kappa(r_d), \kappa(r_t)$:

$$\begin{aligned}\kappa(t_0^*|\hat{c}_0^*, c_1, b_1, t_1) \;=\; & -\kappa(b_1, t_1|c_1)+ \\ \min\Bigg\{ & \kappa(r_b{=}1) + \min\Bigg\{ \begin{array}{l} \kappa(r_d{=}0) + \min\{\kappa(r_t{=}j)|j \in \{4,5,6,7\}\} \\ \kappa(r_d{=}1) + \min\{\kappa(r_t{=}j)|j \in \{1,3,5,7\}\} \\ \kappa(r_d{=}2) + \min\{\kappa(r_t{=}j)|j \in \{4,5,12,13\}\} \\ \kappa(r_d{=}3) + \min\{\kappa(r_t{=}j)|j \in \{1,5,9,13\}\} \end{array} \Bigg\} \\ & \kappa(r_b{=}3) + \min\Bigg\{ \begin{array}{l} \kappa(r_d{=}1) + \min\{\kappa(r_t{=}j)|j \in \{1,3,9,11\}\} \\ \kappa(r_d{=}2) + \min\{\kappa(r_t{=}j)|j \in \{4,6,12,14\}\} \end{array} \Bigg\} \Bigg\}\end{aligned} \tag{4.22}$$

with the following constraints over the response-functions' κ-ranking:

$$\begin{aligned}\min\{\kappa(r_b{=}0), \kappa(r_b{=}1)\} &= \kappa(b_0|c_0) \\ \min\{\kappa(r_b{=}2), \kappa(r_b{=}3)\} &= \kappa(b_1|c_0)\end{aligned}$$

$$\begin{aligned}\min\{\kappa(r_b{=}0), \kappa(r_b{=}2)\} &= \kappa(b_0|c_1) \\ \min\{\kappa(r_b{=}1), \kappa(r_b{=}3)\} &= \kappa(b_1|c_1)\end{aligned}$$

$$\begin{aligned}\min\{\kappa(r_c{=}0), \kappa(r_c{=}1)\} &= \kappa(c_0|c_0) \\ \min\{\kappa(r_c{=}2), \kappa(r_c{=}3)\} &= \kappa(c_1|c_0)\end{aligned}$$

$$\begin{aligned}
\min\{\kappa(r_c{=}0), \kappa(r_c{=}2)\} &= \kappa(c_0|c_1) \\
\min\{\kappa(r_c{=}1), \kappa(r_c{=}3)\} &= \kappa(c_1|c_1)
\end{aligned}$$

$$\begin{aligned}
\min\{\kappa(r_t{=}i)|i \in \{0,1,2,3,4,5,6,7\}\} &= \kappa(t_0|b_0,d_0) \\
\min\{\kappa(r_t{=}i)|i \in \{8,9,10,11,12,13,14,15\}\} &= \kappa(t_1|b_0,d_0)
\end{aligned}$$

$$\begin{aligned}
\min\{\kappa(r_t{=}i)|i \in \{0,1,4,5,8,9,12,13\}\} &= \kappa(t_0|b_0,d_1) \\
\min\{\kappa(r_t{=}i)|i \in \{2,3,6,7,10,11,14,15\}\} &= \kappa(t_1|b_0,d_1)
\end{aligned}$$

$$\begin{aligned}
\min\{\kappa(r_t{=}i)|i \in \{0,1,2,3,8,9,10,11\}\} &= \kappa(t_0|b_1,d_0) \\
\min\{\kappa(r_d{=}i)|i \in \{4,5,6,7,12,13,14,15\}\} &= \kappa(t_1|b_1,d_0)
\end{aligned}$$

$$\begin{aligned}
\min\{\kappa(r_t{=}i)|i \in \{0,2,4,6,8,10,12,14\}\} &= \kappa(t_0|b_1,d_1) \\
\min\{\kappa(r_t{=}i)|i \in \{1,3,5,7,9,11,13,15\}\} &= \kappa(t_1|b_1,d_1)
\end{aligned}$$

We can simplify these constraints by the following procedure. For $\kappa(r_x{=}i)$, find the set of constraints containing $\kappa(r_x{=}i)$ with the maximum right hand side. Then for all other constraints over $\kappa(x|\text{pa}(x))$ eliminate $\kappa(r_x{=}i)$ from each constraint's minimization set. Applying this procedure reduces the above constraints to:

$$\begin{aligned}
\kappa(r_b{=}0) &= 0 \\
\min\{\kappa(r_b{=}2), \kappa(r_b{=}3)\} &= 2 \\
\kappa(r_b{=}1) &= 1
\end{aligned}$$

$$\begin{aligned}
\min\{\kappa(r_d{=}0), \kappa(r_d{=}1)\} &= 3 \\
\kappa(r_d{=}3) &= 0 \\
\kappa(r_d{=}2) &= 1
\end{aligned} \tag{4.23}$$

$$\begin{aligned}
\kappa(r_t{=}6) &= 0 \\
\min\{\kappa(r_t{=}14), \kappa(r_t{=}15)\} &= 2 \\
\min\{\kappa(r_t{=}4), \kappa(r_t{=}5), \kappa(r_t{=}12), \kappa(r_t{=}13)\} &= 3 \\
\min\{\kappa(r_t{=}i)|i \in \{0,1,2,3,8,9,10,11\}\} &= 4 \\
\kappa(r_t{=}7) &= 1
\end{aligned}$$

The conditional kappa term on the right-hand side of Eq. (4.22) may be com-

puted by substituting the conditional kappa rankings specified at the beginning of this section into the following equation:

$$\begin{aligned}\kappa(t_1, b_1|c_1) &= \kappa(b_1|c_1) + \min\left\{\begin{array}{l}\kappa(d_0|c_1) + \kappa(t_1|b_1, d_0)\\ \kappa(d_1|c_1) + \kappa(t_1|b_1, d_1)\end{array}\right\}\\ &= 2\end{aligned}$$

Figure 4.1 represents the structure of Eq. (4.22) and will be used to represent the search state (not the search tree) for finding the upper bound on the κ ranking of the counterfactual. Earlier, we mentioned that each response-function κ value is either constrained to be equal to its lower bound or completely unconstrained. Therefore, each edge in the tree is either assigned to its minimum value or set to ∞. The κ ranking of the counterfactual for these values of the response-function κ's is given by the minimum sum of κ terms over all paths from the root to any leaf node.

To start the search procedure, we assign every response-function κ to its minimum value as given by the right-hand side of Eq. 4.21. This assignment will never violate the constraints imposed by the conditional κ rankings on the observable variables.

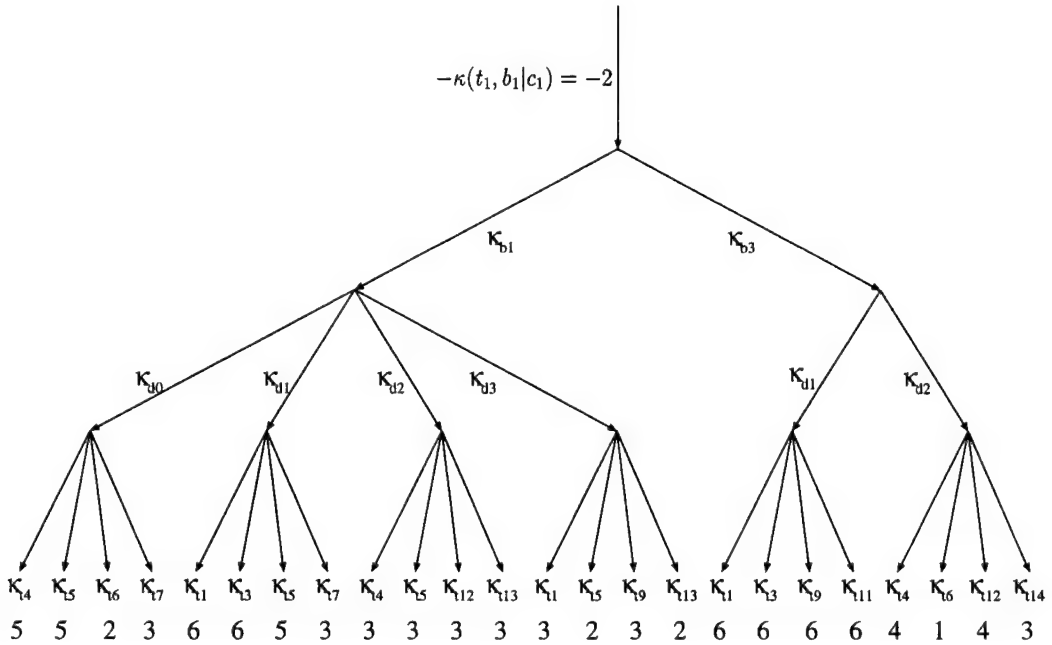

Figure 4.1: *Initial representation of maximization search state.*

We then evaluate the κ sums along each directed path. Taking the minimum

of each sum gives us the lower bound on the κ ranking of the counterfactual, and this also gives us a starting point on the upper bound. The key to the algorithm is that the only way that the upper bound can be greater than the current upper bound, is when response-function κ values may be set to infinite along all directed paths with the next higher κ sums.

Consider, then our current example. Figure 4.2 shows all directed paths with the minimal κ sum equal to 1. In order to extend the bound we must be able to sever this path by assigning one of the response-function κ's to infinite. Those edges along the path that we immediately know cannot be set to infinite because of the simplified constraints are represented with dashed lines. $\kappa(r_b{=}3)$, though, may be set to infinite without violating any constraints. Therefore, we know that the upper bound on the counterfactual's κ ranking is greater than 1.

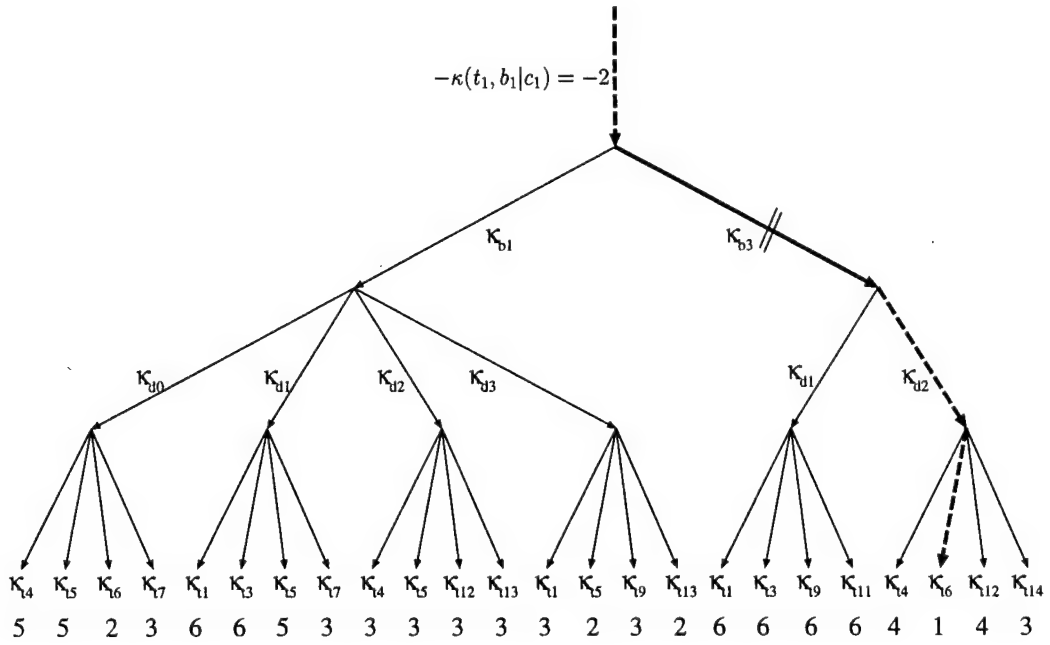

Figure 4.2: *Representation of maximization search state after severing all kappa 1 paths.*

The next step is to consider all directed paths whose κ sums are less than or equal to the next potential upper bound — in this case 2. These paths are shown in Figure 4.3. For this particular example, κ_{d0}, κ_{t5}, and κ_{t13} must also be set to ∞ to sever all paths with κ sums less than or equal to 2. This set of κ's may be set to infinite without violating the given constraints; therefore, the counterfactual κ ranking must be greater than 2.

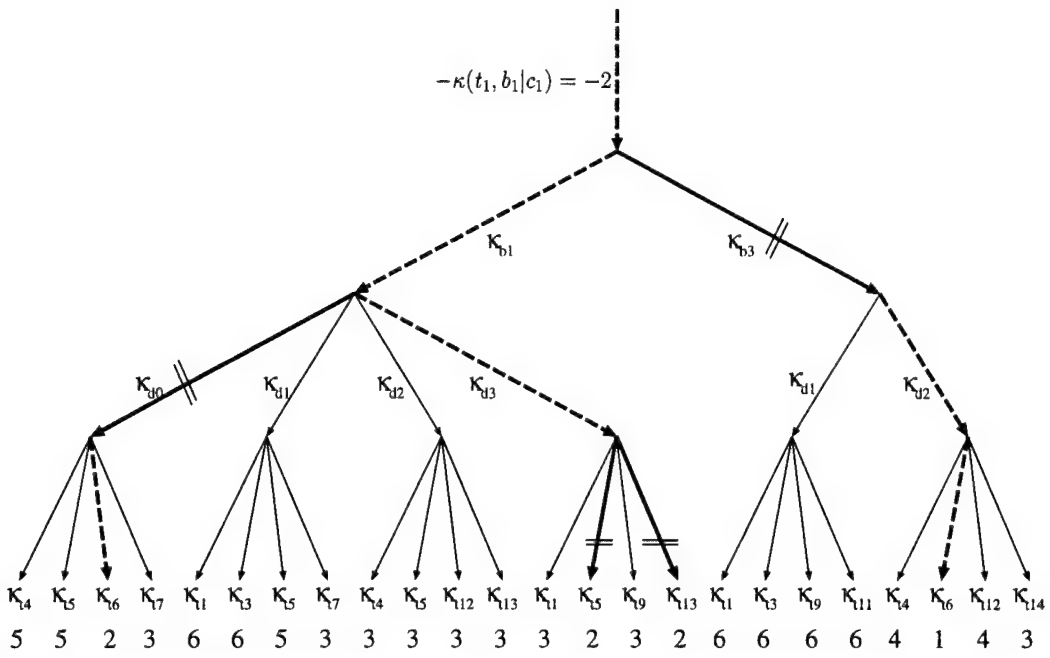

Figure 4.3: *Representation of maximization search state after severing all kappa 2 paths.*

Again we consider all directed paths whose κ sums are less than or equal to the next potential upper bound — now 3. These paths are shown in figure 4.4. In order to sever the left-most three directed paths, both $\kappa(r_d{=}0)$ and $\kappa(r_d{=}1)$ must be set to infinite. However, this violates the constraint given by Eq. (4.23). Therefore, it is impossible to sever all directed paths with κ sums less than or equal to 3, leading us to the conclusion that the upper bound on $\kappa(t_0^*|\hat{c}_0^*, c_1, b_1, t_1)$ is 3. Combined with the earlier results for the lower bound, the range of the counterfactual's κ-ranking is given by

$$1 \leq \kappa(t_0^*|\hat{c}_0^*, c_1, b_1, t_1) \leq 3$$

4.6 Conclusion

In this chapter we have presented a method for evaluating bounds on beliefs in counterfactuals when our general knowledge is given by order-of-magnitude abstractions of probability distributions. Where evaluating bounds on counterfactual probabilities may not succeed because of the presence of local optima in the response-function parameter space, we can always guarantee that the upper

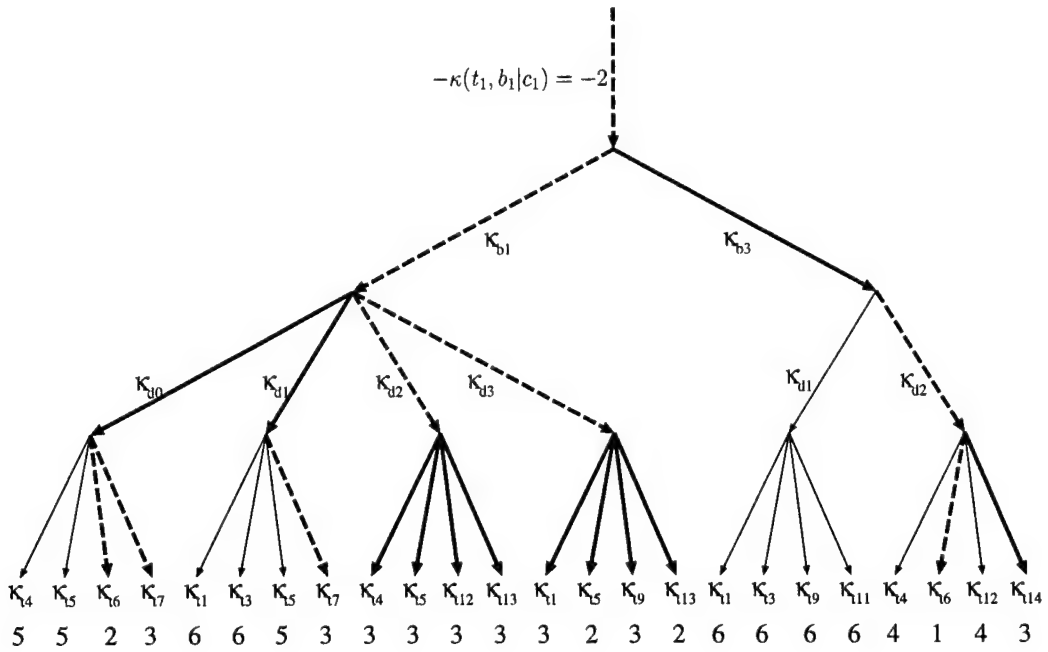

Figure 4.4: *Representation of maximization search state showing that all kappa 3 paths may not be simultaneously severed.*

and lower bounds of the counterfactuals κ ranking may be found given sufficient time. The lower bound on the κ ranking ("we would be at least as surprised as") can be evaluated almost directly once we have the counterfactual's κ rank written in terms of the response-functions' κ rankings. For the upper bound ("we would be at most as surprised as"), an informal algorithm was presented through an example.

Part III

Applications

CHAPTER 5

Clinical trials with imperfect compliance

5.1 Introduction

Consider an experimental study where random assignment has taken place but compliance is not perfect (i.e., the treatment received differs from that assigned). It is well known that under such conditions a bias may be introduced, in the sense that the true causal effect of the treatment may deviate substantially from the causal effect computed by simply comparing subjects receiving the treatment with those not receiving the treatment. Because the subjects who did not comply with the assignment may be precisely those who would have responded adversely (positively) to the treatment, the actual effect of the treatment, when applied uniformly to the population, might be substantially less (more) effective than the study reveals.

In an attempt to avert this bias, economists have devised correctional formulas based on an "instrumental variables" model ([BT84]) which, in general, do not hold outside the linear regression model. A recent analysis by [EF91] departs from the linear regression model, but still makes restrictive commitments to a particular mode of interaction between compliance and response. [Rob89] and [Man90] derived nonparametric bounds on treatment effects using different techniques; however their bounds are not tight. [Hol88] has given a general formulation of the problem (which he called "encouragement design") in terms of Rubin's model of causal effect and has outlined its relation to path analysis and structural equations models. [AIR93], also invoking Rubin's model, have identified a set of assumptions under which the "Instrumental Variable" formula is valid for certain subpopulations. These subpopulations cannot be identified from empirical observation alone, and the need remains to devise alternative, assumption-free formulas for assessing the effect of treatment over the population as a whole. In this chapter, we derive bounds on the average treatment effect that rely solely on observed quantities and are universal, that is, valid no matter what model actually governs the interactions between compliance and response.

The canonical partial-compliance setting can be graphically modeled as shown in Figure 5.1.

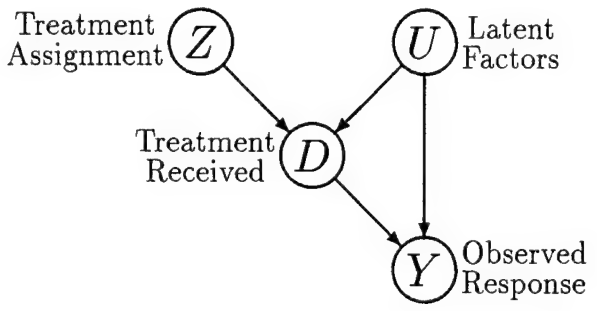

Figure 5.1: *Graphical representation of causal dependencies in a randomized clinical trial with partial compliance.*

We assume that Z, D, and Y are observed binary variables where Z represents the (randomized) treatment assignment, D is the treatment actually received, and Y is the observed response. U represents all factors, both observed and unobserved, that may influence the outcome Y and the treatment D. To facilitate the notation, we let z, d, and y represent, respectively, the values taken by the variables Z, D, and Y, with the following interpretation: $z \in \{z_0, z_1\}$, z_1 asserts that treatment has been assigned (z_0, its negation); $d \in \{d_0, d_1\}$, d_1 asserts that treatment has been administered (d_0, its negation); and $y \in \{y_0, y_1\}$, y_1 asserts a positive observed response (y_0, its negation). The domain of U remains unspecified and may, in general, combine the spaces of several random variables, both discrete and continuous.

The graphical model reflects two assumptions of independence:

1. The treatment assignment does not influence Y directly, but only through the actual treatment D, that is,

$$Z \perp\!\!\!\perp Y \mid \{D, U\} \tag{5.1}$$

   In practice, any direct effect Z might have on Y would be adjusted for through the use of a placebo.

2. Z and U are marginally independent, that is, $Z \perp\!\!\!\perp U$. This independence is partly ensured through the randomization of Z, which rules out a common cause for both Z and U. The absence of a direct path from Z to U represents the assumption that a person's disposition to comply with or deviate from a

given assignment is not in itself affected by the assignment; any such effect can be viewed as part of the disposition.

These assumptions impose on the joint distribution[1] the decomposition

$$P(y,d,z,u) \;=\; P(y|d,u)\; P(d|z,u)\; P(z)\; P(u) \tag{5.2}$$

which, of course, cannot be observed directly because U is a latent variable. However, the marginal distribution $P(y,d,z)$ and, in particular, the conditional distributions $P(y,d|z), z \in \{z_0, z_1\}$, are observed, and the challenge is to assess the causal effect of D on Y from these distributions.[2]

In addition to the independence assumption above, the causal model of Figure 5.1 reflects claims about the behavior of the population under external interventions. In particular, it reflects the assumption that $P(y|d,u)$ is a stable quantity: the probability that an individual with characteristics $U = u$ given treatment $D = d$ will respond with $Y = y$ remains the same, regardless of how the treatment was selected — be it by choice or by policy. Therefore, if we wish to predict the distribution of Y under a condition where the treatment D is applied uniformly to the population, we should calculate

$$P(y^*|\hat{d}^*) \;=\; E_u[P(y|d,u)] \tag{5.3}$$

$$\;=\; \sum_u P(y|d,u)P(u) \tag{5.4}$$

Likewise, if we are interested in estimating the average *change* in Y due to treatment, we define the average *causal effect*, $\mathrm{ACE}(D \to Y)$ ([Hol88]), as

$$\mathrm{ACE}(D \to Y) \;=\; E_u[P(y_1|d_1,u) - P(y_1|d_0,u)] \tag{5.5}$$

$$\;=\; P(y_1^*|\hat{d}_1^*) - P(y_1^*|\hat{d}_0^*) \tag{5.6}$$

For uniformity of notation, we can define, in an analogous way, the average causal effects of the assignment Z, $\mathrm{ACE}(Z \to Y)$ and $\mathrm{ACE}(Z \to D)$. However, since Z is assigned at random, these two quantities can be obtained from the observed distribution:

$$\mathrm{ACE}(Z \to D) \;=\; P(d_1|z_1) - P(d_1|z_0) \tag{5.7}$$

$$\mathrm{ACE}(Z \to Y) \;=\; P(y_1|z_1) - P(y_1|z_0) \tag{5.8}$$

[1] We take the liberty of denoting the prior distribution of U by $P(u)$, even though U may consist of continuous variables.

[2] In practice, of course, only a finite sample of $P(y,d|z)$ will be observed, but since our task is one of identification, not estimation, we make the large-sample assumption and consider $P(y,d|z)$ as given.

The task of causal inference is then to estimate or bound the expression in Eq. (5.6), given the observed probabilities $P(y, d|z_0)$ and $P(y, d|z_1)$.

[Pea93b, Rob89, Man90] have derived bounds on the two terms on the right hand side of Eq. (5.6) given the distribution over Y, D, and Z:

$$\begin{aligned}\max[P(y_1, d_1|z_1);\ & P(y_1, d_1, |z_0)] \\ & \leq E[P(y_1|d_1, u)] \leq \\ & 1 - \max[P(y_0, d_1|z_0);\ P(y_0, d_1|z_1)]\end{aligned} \tag{5.9}$$

$$\begin{aligned}\max[P(y_1, d_0|z_0);\ & P(y_1, d_0, |z_1)] \\ & \leq E[P(y_1|d_0, u)] \leq \\ & 1 - \max[P(y_0, d_0|z_0);\ P(y_0, d_0|z_1)]\end{aligned} \tag{5.10}$$

Choosing appropriate terms to bound the difference

$$E[P(y_1|d_1, u)] - E[P(y_1|d_0, u)]$$

we obtain lower and upper bounds on the causal effect of D on Y:

$$\begin{aligned}P(y_1, d_1|z_1) + P(y_0, d_0|z_0) - 1 & \\ \leq \text{ACE}(D \rightarrow Y) & \leq \\ 1 - P(y_0, d_1|z_1) - P(y_1, d_0|z_0) & \end{aligned} \tag{5.11}$$

or, alternatively,

$$\begin{aligned}\text{ACE}(D \rightarrow Y) &\geq \text{ACE}(Z \rightarrow Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\ \text{ACE}(D \rightarrow Y) &\leq \text{ACE}(Z \rightarrow Y) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0)\end{aligned} \tag{5.12}$$

Due to its simplicity and wide range of applicability, we will call the bounds of Eq. (5.12) the *natural bounds* (three other less intuitive expressions for the upper and lower bounds may be inferred from Eqs. (5.9) and (5.10), but these will not be presented here because they will be derived in Section 5.2). The natural bounds guarantee that the causal effect of the actual treatment cannot exceed that of the intent-to-treat by more than the sum of two measurable quantities, $P(y_1, d_0|z_1) + P(y_0, d_1|z_0)$; they also guarantee that the causal effect of treatment cannot drop below that of the intent-to-treat by more than the sum of two other measurable quantities, $P(y_0, d_0|z_1) + P(y_1, d_1|z_0)$. The width of the natural bound, not surprisingly, is given by the rate of defection, $P(d_1|z_0) + P(d_0|z_1)$.

Before continuing to the more refined derivation of bounds on ACE($D \rightarrow Y$), we should point out that the structural model of Figure 5.1 imposes definite constraints on the observed distributions $P(y, d|z_0)$ and $P(y, d|z_1)$. The constraints, obtained directly from Eq. (5.2), are

$$\begin{aligned} P(y_0, d_1|z_0) + P(y_1, d_1|z_1) &\leq 1 \\ P(y_0, d_1|z_1) + P(y_1, d_1|z_0) &\leq 1 \\ P(y_0, d_0|z_0) + P(y_1, d_0|z_1) &\leq 1 \\ P(y_0, d_0|z_1) + P(y_1, d_0|z_0) &\leq 1 \end{aligned} \tag{5.13}$$

These constraints constitute necessary and sufficient conditions for a marginal probability distribution $P(y, d, z)$ to be generated by the structure of Figure 5.1 (proof in Appendix A.1), and therefore they may serve as an operational test for the compatibility of that structure with the observed data.

5.2 Tight bounds on average causal effect of treatment

Strict bounds on the causal effect of treatment received on subject response may be derived by following the procedure detailed in Section 3.3 where the objective function to be optimized is the difference between the two counterfactual probabilities on the right-hand side of Eq. (5.6).

5.2.1 Response-function model

First, the functional model corresponding to the probabilistic model of Figure 5.1 must be specified. For each of the observable variables in the model (Z, D, and Y), we define the corresponding response-function variables (r_z, r_d, and r_y, respectively).

Figure 5.2 shows the graphical representation of the resulting functional model. Because D and Y are assumed to be influenced by an unobservable common cause, the response-function variables r_d and r_y are connected by an edge.

The states of the variables r_d and r_y have the following interpretations:

D is a deterministic function of the variable Z and $r_d \in \{0, 1, 2, 3\}$:

$$d = f_d(z, r_d) = h_{d,r_d}(z) \tag{5.14}$$

where

$$h_{d,0}(z) = d_0$$

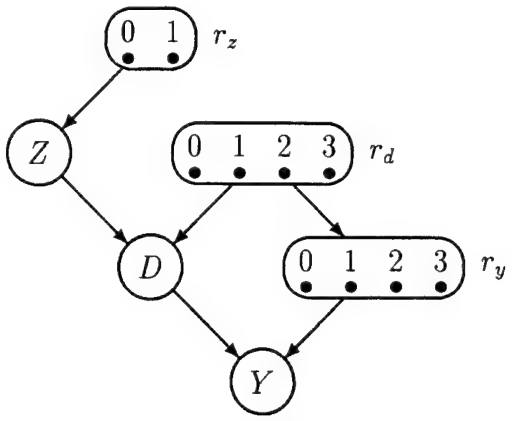

Figure 5.2: *A structure equivalent to that of Figure 5.1 but employing response-function variables r_z, r_d and r_y.*

$$
\begin{aligned}
h_{d,1}(z) &= \begin{cases} d_0 & \text{if } z = z_0 \\ d_1 & \text{if } z = z_1 \end{cases} \\
h_{d,2}(z) &= \begin{cases} d_1 & \text{if } z = z_0 \\ d_0 & \text{if } z = z_1 \end{cases} \\
h_{d,3}(z) &= d_1
\end{aligned}
$$

Similarly, Y is a deterministic function of D and $r_y \in \{0, 1, 2, 3\}$:

$$y = f_y(d, r_y) = h_{y,r_y}(d) \tag{5.15}$$

where

$$
\begin{aligned}
h_{y,0}(d) &= y_0 \\
h_{y,1}(d) &= \begin{cases} y_0 & \text{if } d = d_0 \\ y_1 & \text{if } d = d_1 \end{cases} \\
h_{y,2}(d) &= \begin{cases} y_1 & \text{if } d = d_0 \\ y_0 & \text{if } d = d_1 \end{cases} \\
h_{y,3}(d) &= y_1
\end{aligned}
$$

The correspondence between the states of variables r_d and r_y and the potential response vectors in the Rubin's model [RR83] is rather transparent: each state corresponds to a counterfactual statement specifying how a unit in the population (e.g., a person) would have reacted to any given input. For example, $r_d = 1$ represents units with perfect compliance, while $r_d = 2$ represents units with

perfect defiance. Similarly, $r_y = 1$ represents units with perfect response to treatment, while $r_y = 0$ represents units with no response ($y = y_0$) regardless of treatment. The counterfactual variables Y_1 and Y_0 usually invoked in Rubin's model can be obtained from r_y as follows:

$$Y_1 = \{Y \text{ if } D = d_1\} = \begin{cases} 1 & \text{if } r_y = 1 \text{ or } r_y = 3 \\ 0 & \text{otherwise} \end{cases}$$

$$Y_0 = \{Y \text{ if } D = d_0\} = \begin{cases} 1 & \text{if } r_y = 2 \text{ or } r_y = 3 \\ 0 & \text{otherwise} \end{cases}$$

In general, treatment response and compliance attitudes may not be independent, hence the arrow $r_d \rightarrow r_y$ in Figure 5.2. The joint distribution over $r_d \times r_y$ requires 15 independent parameters, and these parameters are sufficient for specifying the model of Figure 5.2,

$$P(y, d, z, r_d, r_y) = P(y|d, r_y)P(d|r_d, z)P(z)P(r_d, r_y)$$

because Y and D stand in functional relation to their parents in the graph. The causal effect of the treatment can now be obtained directly from Eqs. (5.4) and (5.15) according to Eq. (3.1), giving

$$P(y_1^*|\hat{d}_1^*) = P(r_y{=}1) + P(r_y{=}3) \tag{5.16}$$

$$P(y_1^*|\hat{d}_0^*) = P(r_y{=}2) + P(r_y{=}3) \tag{5.17}$$

and

$$\text{ACE}(D \rightarrow Y) = P(r_y{=}1) - P(r_y{=}2) \tag{5.18}$$

5.2.2 Linear programming formulation

In this section we will explicate the relationship between the parameters of the observed distribution $P(y, d|z)$ and the parameters of the joint distribution $P(r, r')$ of the potential-response functions. This will lead directly to the linear constraints needed for minimizing/maximizing $\text{ACE}(D \rightarrow Y)$ given the observation $P(y, d|z)$.

The conditional distribution $P(y, d|z)$ over the observable variables is fully specified by eight parameters, which will be notated as follows:

$$p_{00.0} = P(y_0, d_0|z_0)$$

$$p_{01.0} = P(y_0, d_1|z_0)$$

$$
\begin{aligned}
p_{10.0} &= P(y_1, d_0|z_0) \\
p_{11.0} &= P(y_1, d_1|z_0)
\end{aligned}
$$

$$
\begin{aligned}
p_{00.1} &= P(y_0, d_0|z_1) \\
p_{01.1} &= P(y_0, d_1|z_1) \\
p_{10.1} &= P(y_1, d_0|z_1) \\
p_{11.1} &= P(y_1, d_1|z_1)
\end{aligned}
$$

The probabilistic constraints

$$\sum_{n=00}^{11} p_{n.0} = 1 \tag{5.19}$$

$$\sum_{n=00}^{11} p_{n.1} = 1 \tag{5.20}$$

further imply that $\vec{p} = (p_{00.0}, p_{01.0}, p_{10.0}, p_{11.0}, p_{00.1}, p_{01.1}, p_{10.1}, p_{11.1})$ can be specified by a point in six-dimensional space. This space will be referred to as P. Eqs. (5.7) and (5.8) may be rewritten in terms of these parameters as

$$\text{ACE}(Z \to D) = p_{11.1} + p_{01.1} - p_{11.0} - p_{01.0} \tag{5.21}$$

$$\text{ACE}(Z \to Y) = p_{11.1} + p_{10.1} - p_{11.0} - p_{10.0} \tag{5.22}$$

The joint probability over $R \times R'$, $P(r, r')$, has 16 parameters and completely specifies the population under study. These parameters will be notated as

$$q_{jk} = P(r = r_j, r' = r'_k)$$

where $j, k \in \{0, 1, 2, 3\}$. The probabilistic constraint

$$\sum_{j=0}^{3} \sum_{k=0}^{3} q_{jk} = 1$$

implies that $\vec{q} = (q_{00}, q_{01}, q_{02}, q_{03}, q_{10}, q_{11}, q_{12}, q_{13}, q_{20}, q_{21}, q_{22}, q_{23}, q_{30}, q_{31}, q_{32}, q_{33})$ specifies a point in 15-dimensional space. This space will be referred to as Q.

Eq. (5.18) can now be rewritten as a linear combination of the Q parameters:

$$\text{ACE}(D \to Y) = q_{01} + q_{11} + q_{21} + q_{31} - q_{02} - q_{12} - q_{22} - q_{32} \tag{5.23}$$

Given some point $\vec{q}$ in Q space, there is a direct linear transformation to the corresponding point $\vec{p}$ in the observation space P:

$$\begin{aligned} p_{00.0} &= q_{00} + q_{01} + q_{10} + q_{11} \\ p_{01.0} &= q_{20} + q_{22} + q_{30} + q_{32} \\ p_{10.0} &= q_{02} + q_{03} + q_{12} + q_{13} \\ p_{11.0} &= q_{21} + q_{23} + q_{31} + q_{33} \end{aligned} \tag{5.24}$$

$$\begin{aligned} p_{00.1} &= q_{00} + q_{01} + q_{20} + q_{21} \\ p_{01.1} &= q_{10} + q_{12} + q_{30} + q_{32} \\ p_{10.1} &= q_{02} + q_{03} + q_{22} + q_{23} \\ p_{11.1} &= q_{11} + q_{13} + q_{31} + q_{33} \end{aligned} \tag{5.25}$$

which will sometimes be written in matrix form, $\vec{p} = \bar{P}\vec{q}$.

Given a point $\vec{p}$ in P space, the strict lower bound on $\mathrm{ACE}(D \rightarrow Y)$ can be determined by solving the following linear programming problem:

Minimize: $q_{01} + q_{11} + q_{21} + q_{31} - q_{02} - q_{12} - q_{22} - q_{32}$

Subject to:

$$\begin{aligned} \sum_{j=0}^{3}\sum_{k=0}^{3} q_{jk} &= 1 \\ \bar{P}\vec{q} &= \vec{p} \\ q_{jk} &\geq 0 \text{ for } j,k \in \{0,1,2,3\} \end{aligned} \tag{5.26}$$

5.3 Closed-form solutions to the linear programming problem

Given an observed point $\vec{p}$ in P space, $L_{D\rightarrow Y}(\vec{p})$ and $U_{D\rightarrow Y}(\vec{p})$, respectively, will represent the strict lower and upper bounds on $\mathrm{ACE}(D \rightarrow Y)$ associated with $\vec{p}$. More precisely,

$$L_{D\rightarrow Y}(\vec{p}) = \min_{\vec{q} \text{ s.t. } \vec{p}=\bar{P}\vec{q}} \mathrm{ACE}(D \rightarrow Y) \tag{5.27}$$

$$U_{D\rightarrow Y}(\vec{p}) = \max_{\vec{q} \text{ s.t. } \vec{p}=\bar{P}\vec{q}} \mathrm{ACE}(D \rightarrow Y) \tag{5.28}$$

where Eq. (5.23) gives $\mathrm{ACE}(D \rightarrow Y)$ in terms of $\vec{q}$.

For every given point $\vec{p}$, the optimization above can be executed using the Simplex Tableau algorithm (see [DM81]), which yields a pair of numerical values for $L_{D\to Y}(\vec{p})$ and $U_{D\to Y}(\vec{p})$. Fortunately, the size of the problem permits a closed-form solution to be obtained by enumerating all vertices of the dual linear-programming problem's constraint polygon (see Appendix B). This procedure leads to the following bounds:

$$L_{D\to Y}(\vec{p}) = \max\left\{\begin{array}{c} p_{11.1}+p_{00.0}-1 \\ p_{11.0}+p_{00.1}-1 \\ -p_{01.1}-p_{10.1} \\ -p_{01.0}-p_{10.0} \\ p_{11.0}-p_{11.1}-p_{10.1}-p_{01.0}-p_{10.0} \\ p_{11.1}-p_{11.0}-p_{10.0}-p_{01.1}-p_{10.1} \\ p_{00.1}-p_{01.1}-p_{10.1}-p_{01.0}-p_{00.0} \\ p_{00.0}-p_{01.0}-p_{10.0}-p_{01.1}-p_{00.1} \end{array}\right\} \tag{5.29}$$

$$U_{D\to Y}(\vec{p}) = \min\left\{\begin{array}{c} 1-p_{01.1}-p_{10.0} \\ 1-p_{01.0}-p_{10.1} \\ p_{11.1}+p_{00.1} \\ p_{11.0}+p_{00.0} \\ -p_{01.0}+p_{01.1}+p_{00.1}+p_{11.0}+p_{00.0} \\ -p_{01.1}+p_{11.1}+p_{00.1}+p_{01.0}+p_{00.0} \\ -p_{10.1}+p_{11.1}+p_{00.1}+p_{11.0}+p_{10.0} \\ -p_{10.0}+p_{11.0}+p_{00.0}+p_{11.1}+p_{10.1} \end{array}\right\} \tag{5.30}$$

Note that the first term in these two expressions correspond to the natural bounds of Eq. (5.11). Tables 5.1 and 5.2 list the regions of P space for which each of the terms in Eqs. (5.29) and (5.30) represents the lower/upper bound, respectively. These bounds constitute substantial improvement over those derived by Robins (1989) and Manski (1990), which correspond to the four upper terms in both (5.29) and (5.30). The width of these bounds cannot exceed the rate of noncompliance, $P(d_1|z_0)+P(d_0|z_1)$.

We may also derive bounds on the treatment responses under the condition that one treatment is uniformly applied to the population, by optimizing Eqs. (5.16) and (5.17) individually (under the same linear constraints). The

Conditions	$L_{D\to Y}(\vec{p})$
$p_{11.1} \geq p_{11.0}$ $p_{01.1} + p_{10.1} \geq p_{01.0}$ $p_{00.0} \geq p_{00.1}$ $p_{01.0} + p_{10.0} \geq p_{10.1}$	$p_{11.1} + p_{00.0} - 1$
$p_{11.0} \geq p_{11.1}$ $p_{01.0} + p_{10.0} \geq p_{01.1}$ $p_{00.1} \geq p_{00.0}$ $p_{01.1} + p_{10.1} \geq p_{10.0}$	$p_{11.0} + p_{00.1} - 1$
$p_{11.0} + p_{10.0} \geq p_{11.1} \geq p_{11.0}$ $p_{01.0} + p_{00.0} \geq p_{00.1} \geq p_{00.0}$	$-p_{01.1} - p_{10.1}$
$p_{11.1} + p_{10.1} \geq p_{11.0} \geq p_{11.1}$ $p_{01.1} + p_{00.1} \geq p_{00.0} \geq p_{00.1}$	$-p_{01.0} - p_{10.0}$
$p_{11.0} \geq p_{11.1} + p_{10.1}$ $p_{01.1} \geq p_{01.0} + p_{10.0}$	$p_{11.0} - p_{11.1} - p_{10.1} - p_{01.0} - p_{10.0}$
$p_{11.1} \geq p_{11.0} + p_{10.0}$ $p_{01.0} \geq p_{01.1} + p_{10.1}$	$p_{11.1} - p_{01.1} - p_{10.1} - p_{11.0} - p_{10.0}$
$p_{10.0} \geq p_{01.1} + p_{10.1}$ $p_{00.1} \geq p_{01.0} + p_{00.0}$	$p_{00.1} - p_{01.1} - p_{10.1} - p_{01.0} - p_{00.0}$
$p_{10.1} \geq p_{01.0} + p_{10.0}$ $p_{00.0} \geq p_{01.1} + p_{00.1}$	$p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1}$

Table 5.1: *Lower bounds on* $\mathrm{ACE}(D \to Y)$ *given a point* $\vec{p}$ *in the observation space* P.

Conditions	$U_{D\to Y}(\vec{p})$
$p_{01.1} \geq p_{01.0}$ $p_{11.1} + p_{00.1} \geq p_{11.0}$ $p_{10.0} \geq p_{10.1}$ $p_{11.0} + p_{00.0} \geq p_{00.1}$	$1 - p_{01.1} - p_{10.0}$
$p_{01.0} \geq p_{01.1}$ $p_{11.0} + p_{00.0} \geq p_{11.1}$ $p_{10.1} \geq p_{10.0}$ $p_{11.1} + p_{00.1} \geq p_{00.0}$	$1 - p_{01.0} - p_{10.1}$
$p_{01.0} + p_{00.0} \geq p_{01.1} \geq p_{01.0}$ $p_{11.0} + p_{10.0} \geq p_{10.1} \geq p_{10.0}$	$p_{11.1} + p_{00.1}$
$p_{01.1} + p_{00.1} \geq p_{01.0} \geq p_{01.1}$ $p_{11.1} + p_{10.1} \geq p_{10.0} \geq p_{10.1}$	$p_{11.0} + p_{00.0}$
$p_{01.0} \geq p_{01.1} + p_{00.1}$ $p_{11.1} \geq p_{11.0} + p_{00.0}$	$-p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0}$
$p_{01.1} \geq p_{01.0} + p_{00.0}$ $p_{11.0} \geq p_{11.1} + p_{00.1}$	$-p_{01.1} + p_{11.1} + p_{00.1} + p_{01.0} + p_{00.0}$
$p_{00.0} \geq p_{11.1} + p_{00.1}$ $p_{10.1} \geq p_{11.0} + p_{10.0}$	$-p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0}$
$p_{00.1} \geq p_{11.0} + p_{00.0}$ $p_{10.0} \geq p_{11.1} + p_{10.1}$	$-p_{10.0} + p_{11.0} + p_{00.0} + p_{11.1} + p_{10.1}$

Table 5.2: *Upper bounds on* $\mathrm{ACE}(D \to Y)$ *given a point* $\vec{p}$ *in observation space* P.

resulting bounds are:

$$\max \left\{ \begin{array}{c} p_{10.0} + p_{11.0} - p_{00.1} - p_{11.1} \\ p_{10.1} \\ p_{10.0} \\ p_{01.0} + p_{10.0} - p_{00.1} - p_{01.1} \end{array} \right\}$$
$$\leq P(y_1^* | \hat{d}_0^*) \leq$$
$$\min \left\{ \begin{array}{c} p_{01.0} + p_{10.0} + p_{10.1} + p_{11.1} \\ 1 - p_{00.1} \\ 1 - p_{00.0} \\ p_{10.0} + p_{11.0} + p_{01.1} + p_{10.1} \end{array} \right\}$$

and

$$\max \left\{ \begin{array}{c} p_{11.0} \\ p_{11.1} \\ -p_{00.0} - p_{01.0} + p_{00.1} + p_{11.1} \\ -p_{01.0} - p_{10.0} + p_{10.1} + p_{11.1} \end{array} \right\}$$
$$\leq P(y_1^* | \hat{d}_1^*) \leq$$
$$\min \left\{ \begin{array}{c} 1 - p_{01.1} \\ 1 - p_{01.0} \\ p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} \\ p_{10.0} + p_{11.0} + p_{00.1} + p_{11.1} \end{array} \right\}$$

These bounds improve upon the results of [Man90]. In addition, one can prove that these are the tightest possible assumption-free bounds.

5.3.1 The positive-effects convention

To simplify the presentation of the bounds found in the last subsection, we first choose a notational system in which assignment to treatment does not reduce the probability of treatment usage ($D = d_1$) and of positive response ($Y = y_1$). From Eqs. (5.7) and (5.8), these conditions can be written as

$$\begin{aligned} \text{ACE}(Z \rightarrow D) &\geq 0 \\ \text{ACE}(Z \rightarrow Y) &\geq 0 \end{aligned}$$

or, alternatively,

$$\begin{aligned} p_{01.1} + p_{11.1} &\geq p_{01.0} + p_{11.0} \\ p_{10.1} + p_{11.1} &\geq p_{10.0} + p_{11.0} \end{aligned}$$

The conjunction of these two inequalities will be referred to as the *condition of positive effects.* This constraint may be imposed without loss of generality, because the labels of the variables' values can always be swapped in such a way that the inequalities are satisfied: if $\text{ACE}(Z \to D) < 0$, we swap d_0 and d_1; if $\text{ACE}(Z \to Y) < 0$, we swap y_0 and y_1.

In a notational system where the condition of positive effects holds, the lower and upper bounds on the treatment effect can be simplified to read

$$L_{D\to Y}(\vec{p}) = \max \left\{ \begin{array}{c} p_{11.1} + p_{00.0} - 1 \\ p_{11.1} - p_{11.0} - p_{10.0} - p_{01.1} - p_{10.1} \\ -p_{01.1} - p_{10.1} \\ -p_{01.0} - p_{10.0} \\ p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1} \end{array} \right\} \tag{5.31}$$

and

$$U_{D\to Y}(\vec{p}) = \min \left\{ \begin{array}{c} 1 - p_{01.1} - p_{10.0} \\ 1 - p_{01.0} - p_{10.1} \\ -p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0} \\ p_{11.1} + p_{00.1} \\ p_{11.0} + p_{00.0} \\ -p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0} \end{array} \right\} \tag{5.32}$$

respectively.

5.3.2 Graphical presentation of the bounds

When compliance is perfect (i.e., $\text{ACE}(Z \to D) = 1$), we expect the causal effect of the treatment to coincide with the causal effect of the intent-to-treat, that is,

$$\text{ACE}(D \to Y) = \text{ACE}(Z \to Y) \quad \text{if} \quad \text{ACE}(Z \to D) = 1$$

Similarly, if all units were to exhibit the same difference in compliance probabilities, $P(d_1|z_1, u) - P(d_1|z_0, u)$, the celebrated "Instrumental Variable" formula applies

$$\text{ACE}(D \to Y) = \frac{\text{ACE}(Z \to Y)}{\text{ACE}(Z \to D)} = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1|z_1) - P(d_1|z_0)} \tag{5.33}$$

Here $\text{ACE}(D \to Y)$ is determined solely by $\text{ACE}(Z \to Y)$ and $\text{ACE}(Z \to D)$. In general, however, the latter two parameters will not be sufficient to determine $\text{ACE}(D \to Y)$ uniquely; nevertheless, they can be used to determine the range within which $\text{ACE}(D \to Y)$ may fall.

Figure 5.3 plots $L_{D\to Y}(\vec{p})$ and $U_{D\to Y}(\vec{p})$ given $\mathrm{ACE}(Z \to D)$ and $\mathrm{ACE}(Z \to Y)$. The range of $\mathrm{ACE}(D \to Y)$ is quite wide, and is given by the simple formula:

$$\begin{aligned}\mathrm{ACE}(Z \to Y) + \mathrm{ACE}(Z \to D) - 1 \\ \leq \mathrm{ACE}(D \to Y) \leq \\ 1 - |\mathrm{ACE}(Z \to D) - \mathrm{ACE}(Z \to Y)|\end{aligned} \tag{5.34}$$

An interesting point is that plotting the natural bounds given by Eq. (5.12) as a function of $\mathrm{ACE}(Z \to D)$ and $\mathrm{ACE}(Z \to Y)$ gives us precisely the same results as shown in Figure 5.3.

Note that the bounds $L_{D\to Y}(\vec{p})$ and $U_{D\to Y}(\vec{p})$ for a particular point $\vec{p}$ in P space may be much tighter than the bounds shown in Figure 5.3 as functions of $\mathrm{ACE}(Z \to D)$ and $\mathrm{ACE}(Z \to Y)$ evaluated at $\vec{p}$. This will be demonstrated by example in Section 5.4.

5.4 Examples

At this point it is worth summarizing by example how the bounds of Eqs. (5.29) and (5.30) can be used to provide meaningful information about causal effects.

Consider the Lipid Research Clinics Coronary Primary Prevention Trial data (see [Pro84] for an extended description of the clinical trial). A portion of this data consisting of 337 subjects was analyzed in [EF91] using a model that incorporated subject compliance as an explanatory variable; this same data set is the focus of our analysis. A population of subjects was assembled and two preliminary cholesterol measurements were obtained: one prior to a suggested low-cholesterol diet (continuous variable C_{I1}); and one following the diet period (C_{I2}). The initial cholesterol level (C_I) was taken as a weighted average of these two measures: $C_I = 0.25C_{I1} + 0.75C_{I2}$. The subjects were randomized into two treatment groups; in the first group all subjects were prescribed cholestyramine (z_1), while the subjects in the other group were prescribed a placebo (z_0). During several years of treatment, each subject's cholesterol level was measured multiple times, and the average of these measurements was used as the post-treatment cholesterol level (continuous variable C_F). The compliance of each subject was determined by tracking the quantity of prescribed dosage consumed (continuous variable B).

In order to apply our analysis to this study, the continuous data obtained

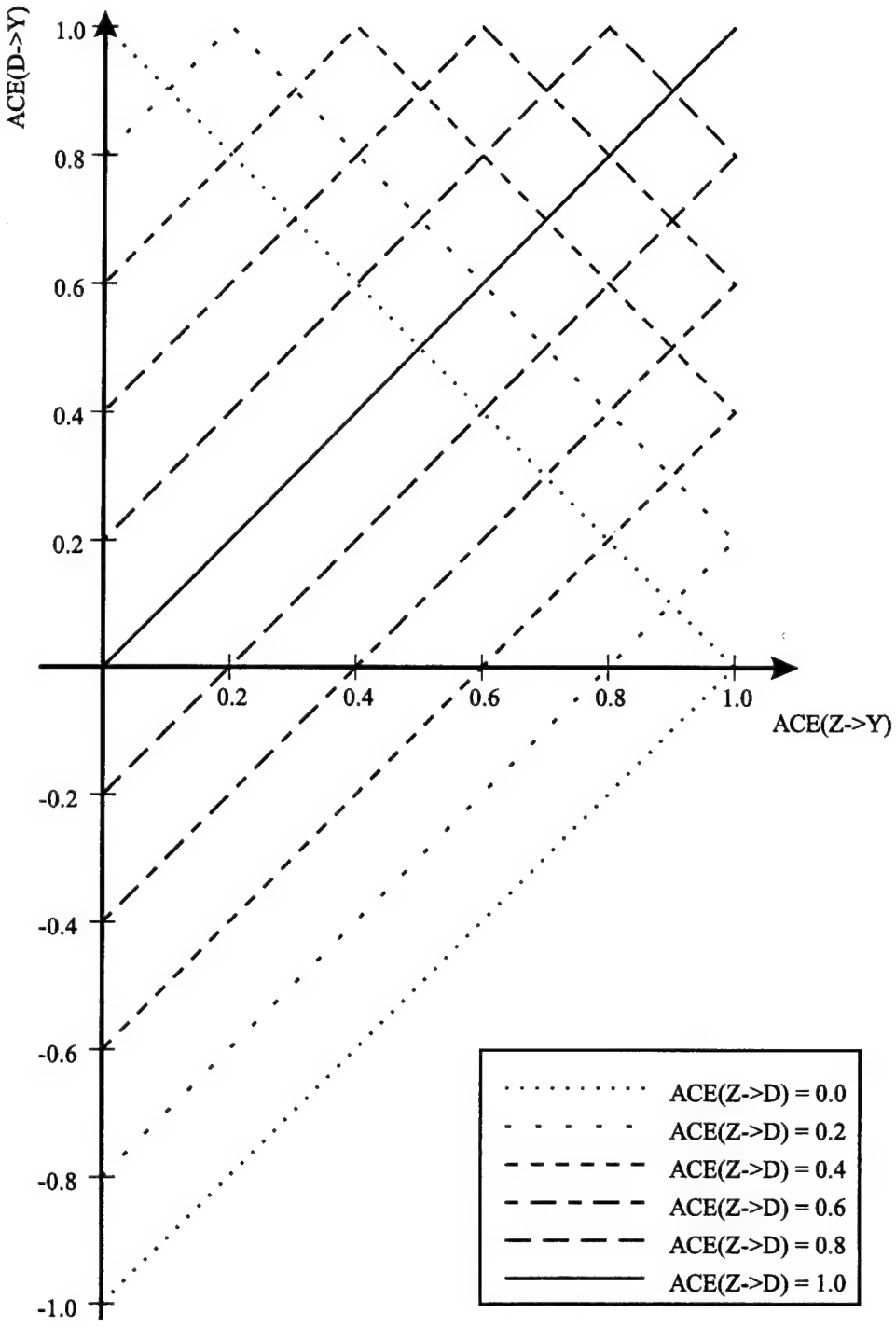

Figure 5.3: *Bounds on* $\mathrm{ACE}(D \to Y)$ *plotted against* $\mathrm{ACE}(Z \to Y)$ *and* $\mathrm{ACE}(Z \to D)$.

in the [Pro84] study must be transformed to binary variables representing *treatment assignment* (Z), *received treatment* (D), and *treatment response* (Y). The following transformation accomplishes this by thresholding dosage consumption and change in cholesterol level:

$$d = \begin{cases} d_0 & \text{if } z = z_0 \text{ or } b < 50 \\ d_1 & \text{if } z = z_1 \text{ and } b \geq 50 \end{cases} \tag{5.35}$$

$$y = \begin{cases} y_0 & \text{if } c_I - c_F < 28 \\ y_1 & \text{if } c_I - c_F \geq 28 \end{cases} \tag{5.36}$$

This transformation reflects the assumption that a subject does not receive cholestyramine if not assigned to the cholestyramine treatment group, namely, $P(y_0, d_1|z_0) = 0$ and $P(y_1, d_1|z_0) = 0$. The threshold for dosage consumption in Eq. (5.35) was selected as roughly the midpoint between minimum and maximum consumption, while the threshold for cholesterol level reduction in Eq. (5.36) was selected at 28 units.

If the data samples are interpreted according to Eqs. (5.35) and (5.36), then the computed distribution over (Z, D, Y) results in the following point in P space[3]:

$$\begin{aligned} p_{00.0} = P(y_0, d_0|z_0) &= 0.919 \\ p_{01.0} = P(y_0, d_1|z_0) &= 0.000 \\ p_{10.0} = P(y_1, d_0|z_0) &= 0.081 \\ p_{11.0} = P(y_1, d_1|z_0) &= 0.000 \end{aligned}$$

$$\begin{aligned} p_{00.1} = P(y_0, d_0|z_1) &= 0.315 \\ p_{01.1} = P(y_0, d_1|z_1) &= 0.139 \\ p_{10.1} = P(y_1, d_0|z_1) &= 0.073 \\ p_{11.1} = P(y_1, d_1|z_1) &= 0.473 \end{aligned}$$

By first computing the causal effects of the intent-to-treat,

$$\begin{aligned} \text{ACE}(Z \rightarrow D) &= p_{11.1} + p_{01.1} - p_{11.0} - p_{01.0} = 0.612 \\ \text{ACE}(Z \rightarrow Y) &= p_{11.1} + p_{10.1} - p_{11.0} - p_{10.0} = 0.465 \end{aligned} \tag{5.37}$$

[3]We make the large-sample assumption and take the sample frequencies as representing $P(y, d|z)$.

we can verify that the condition of positive effects is satisfied. This justifies the use of Eqs. (5.31) and (5.32) for evaluating the strict lower and upper bounds on $\text{ACE}(D \rightarrow Y)$. By computing the quantities required for Eq. (5.31), we obtain

$$L_{D\rightarrow Y}(\vec{p}) = \max \left\{ \begin{array}{rcr} p_{11.1} + p_{00.0} - 1 & = & 0.392 \\ p_{11.1} - p_{11.0} - p_{10.0} - p_{01.1} - p_{10.1} & = & 0.180 \\ -p_{01.1} - p_{10.1} & = & -0.212 \\ -p_{01.0} - p_{10.0} & = & -0.081 \\ p_{00.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{00.1} & = & 0.384 \end{array} \right\}$$

Those needed for Eq. (5.32) give us

$$U_{D\rightarrow Y}(\vec{p}) = \min \left\{ \begin{array}{rcr} 1 - p_{01.1} - p_{10.0} & = & 0.780 \\ 1 - p_{01.0} - p_{10.1} & = & 0.927 \\ -p_{01.0} + p_{01.1} + p_{00.1} + p_{11.0} + p_{00.0} & = & 1.373 \\ p_{11.1} + p_{00.1} & = & 0.788 \\ p_{11.0} + p_{00.0} & = & 0.919 \\ -p_{10.1} + p_{11.1} + p_{00.1} + p_{11.0} + p_{10.0} & = & 0.796 \end{array} \right\}$$

Accordingly, we conclude that the treatment causal effect lies in the range

$$0.392 \leq \text{ACE}(D \rightarrow Y) \leq 0.780 \tag{5.38}$$

which is rather remarkable; the experimenter can categorically state that when applied uniformly to the population, the treatment is guaranteed to improve by at least 39.2% the probability of reducing the level of cholesterol by at least 28 points. This guarantee does not rest on any assumed model. Unfortunately, these results cannot be translated directly into a useful policy statement for treating people with high cholesterol, because the [Pro84] data were obtained for continuous level of dosage consumed (D), while our analysis is restricted to binary D. To infer the behavior of the population under uniform consumption at a specific level of dosage, a model with a continuous (or at least 3-level) treatment must be studied; these types of models will be addressed in Chapter 6.

Note that the bounds in Eq. (5.38) are equal to the natural bounds given by Eq. (5.12):

$$\begin{aligned} \text{ACE}(D \rightarrow Y) &\geq 0.465 - 0.073 - 0.000 = 0.392 \\ \text{ACE}(D \rightarrow Y) &\leq 0.465 + 0.315 + 0.000 = 0.780 \end{aligned}$$

It is interesting to note that "naive" comparison of subjects in and out of the treatment group would predict, in this case, the value of

$$P(y_1|d_1) - P(y_1|d_0) = 0.662$$

which demonstrates the potential inaccuracy in using the mean difference for evaluating ACE($D \rightarrow Y$).

If ACE($Z \rightarrow D$) and ACE($Z \rightarrow Y$) are the only quantities measured, then the following bounds on ACE($D \rightarrow Y$) can be computed by substituting the values from Eq. (5.37) into Eq. (5.34):

$$0.077 \leq \text{ACE}(D \rightarrow Y) \leq 0.853$$

As noted in Section 5.3.2, these bounds are much wider than those obtained in Eq. (5.38), which utilized the full information given by $P(y, d|z)$.

5.5 Tightness of the natural bound

Although the example above shows no improvement over the natural bounds, the next (hypothetical) example will show that in certain cases the natural bounds can be improved upon significantly. Consider the following point in P space:

$$\begin{aligned}
p_{00.0} = P(y_0, d_0|z_0) &= 0.55 \\
p_{01.0} = P(y_0, d_1|z_0) &= 0.45 \\
p_{10.0} = P(y_1, d_0|z_0) &= 0.00 \\
p_{11.0} = P(y_1, d_1|z_0) &= 0.00
\end{aligned}$$

$$\begin{aligned}
p_{00.1} = P(y_0, d_0|z_1) &= 0.45 \\
p_{01.1} = P(y_0, d_1|z_1) &= 0.00 \\
p_{10.1} = P(y_1, d_0|z_1) &= 0.00 \\
p_{11.1} = P(y_1, d_1|z_1) &= 0.55
\end{aligned}$$

Substitution of these parameters into Eq. (5.12) results in the natural bounds

$$0.10 \leq \text{ACE}(D \rightarrow Y) \leq 0.55$$

while the bounds resulting from the application of Eqs. (5.29) and (5.30) collapse to

$$0.55 \leq \text{ACE}(D \rightarrow Y) \leq 0.55$$

Obviously, when our goal is the assessment of the treatment causal effect, the bounds obtained through linear programming can be much more informative.

Interestingly, a precise determination of ACE($D \rightarrow Y$) is feasible even though the compliance is low:

$$\text{ACE}(Z \rightarrow D) \;=\; 0.10$$

Intuitively, one would expect that if most subjects ignore their treatment assignment, the results of the study would be suspect. This intuition is partially supported by Figure 5.3, which shows that the feasible range of ACE($D \rightarrow Y$) tends to widen as ACE($Z \rightarrow D$) decreases. Nevertheless, the idiosyncratic features of the data in this example permit us to determine precisely the causal effect. These features also allow us to precisely determine the distribution of subjects in the population, in terms of the subjects' compliance and response characteristics.

The first behavior is characterized by perfect compliance with the assignment along with a perfect response pattern to the treatment received ($y = y_1$ if and only if $d = d_1$). The second behavior is characterized by perfect defiance of the assignment (the subject always chooses the treatment that is the opposite of the one assigned) along with a total inability to respond positively to either treatment. The strong and strange interactions between the compliance and response behaviors implied by these data would be very uncharacteristic of most subject populations.

In fact, we can prove that there are exactly six regions where the average causal effect of treatment on response is identifiable when no assumptions are presumed. This is accomplished by enumerating the conditions whereby one of the lower bound terms in Eq. (5.29) is equal to one of the upper bound terms in Eq. (5.30).

Region	ACE($D \rightarrow Y$)
$P(d_1\|z_0) = 0$ $P(d_0\|z_1) = 0$	$P(y_1, d_1\|z_1) + P(y_0, d_0\|z_0) - 1$
$P(y_1, d_1\|z_0) = 0$ $P(y_0, d_1\|z_1) = 0$ $P(y_0, d_1\|z_0) + P(y_1, d_1\|z_1) = 1$	$P(y_0, d_0\|z_0) + p(y_0, d_0\|z_1) - P(y_0, d_1\|z_0)$
$P(y_1, d_0\|z_0) = 0$ $P(y_0, d_0\|z_1) = 0$ $P(y_0, d_0\|z_0) + P(y_1, d_0\|z_1) = 1$	$P(y_1, d_1\|z_1) + P(y_1, d_1\|z_0) - P(y_1, d_0\|z_1)$
$P(d_0\|z_0) = 0$ $P(d_1\|z_1) = 0$	$P(y_1, d_1\|z_0) + P(y_0, d_0\|z_1) - 1$
$P(y_0, d_1\|z_0) = 0$ $P(y_1, d_1\|z_1) = 0$ $P(y_1, d_1\|z_0) + P(y_0, d_1\|z_1) = 1$	$P(y_0, d_0\|z_0) + p(y_0, d_0\|z_1) - P(y_0, d_1\|z_1)$
$P(y_0, d_0\|z_0) = 0$ $P(y_1, d_0\|z_1) = 0$ $P(y_1, d_0\|z_0) + P(y_0, d_0\|z_1) = 1$	$P(y_1, d_1\|z_1) + P(y_1, d_1\|z_0) - P(y_1, d_0\|z_0)$

The entries in this table indicate that precise determination of treatment effects is feasible whenever (a) the percentage of subjects complying with assignment z_0 is the same as those complying with z_1 and (b) in at least one treatment arm d, y and z are perfectly correlated.

In this section, we have shown that, in general, the natural bounds given by Eq. (5.12) may not always be as tight as the bounds given by Eqs. (5.29) and (5.30). In the next section, however, we will demonstrate that the natural bounds are tight in two important subspaces of P: when the data reveal treatment sufficiency (conditional independence between treatment assignment and treatment response given treatment received), and when it is reasonable to assume that subjects are *non-defiant.*

5.6 Incorporating additional assumptions

In this section we will examine the impact that various assumptions have on the bounds for ACE($D \rightarrow Y$) and the constraints that they place on the observed parameters. The main assumptions to be discussed here are:

- treatment sufficiency (conditional independence of treatment assignment and observed response given treatment received);

- treatment sufficiency together with structural stability;
- no perfectly defiant subjects; and
- monotonic compliance and response behaviors.

5.6.1 Treatment sufficiency

This subsection examines whether the presence of conditional independence $Z \perp\!\!\!\perp Y|D$ in the data simplifies the formulas for the bounds on $\mathrm{ACE}(D \rightarrow Y)$. In other words, are any of the expressions within the minimization/maximization of Eqs. (5.31) and (5.32) eliminated? The following theorem provides the answer to this question.

Theorem 5.6.1 *If the observed distribution $P(y,d|z)$ satisfies $Z \perp\!\!\!\perp Y|D$ and the condition of positive effects, then the natural bounds on* $\mathrm{ACE}(D \rightarrow Y)$

$$\begin{aligned} \mathrm{ACE}(D \rightarrow Y) &\geq \mathrm{ACE}(Z \rightarrow Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\ \mathrm{ACE}(D \rightarrow Y) &\leq \mathrm{ACE}(Z \rightarrow Y) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0) \end{aligned}$$

are tight.

Proof:

We will show that a set of constraints implied by $Z \perp\!\!\!\perp Y|D$ and the condition of positive effects are only mutually consistent with those conditions in Tables 5.1 and 5.2 corresponding to the natural bounds (the topmost entries).

First, assume that $\vec{p}$ is strictly positive.

By definition, $Z \perp\!\!\!\perp Y|D$ if and only if

$$P(y|d, z_0) = P(y|d, z_1)$$

for all y and d such that $P(d|z_0) > 0$ and $P(d|z_1) > 0$. This may be written:

$$\begin{aligned} \frac{p_{10.0}}{p_{00.0} + p_{10.0}} &= \frac{p_{10.1}}{p_{00.1} + p_{10.1}} \\ \frac{p_{11.0}}{p_{01.0} + p_{11.0}} &= \frac{p_{11.1}}{p_{01.1} + p_{11.1}} \end{aligned}$$

or, equivalently,

$$\begin{aligned} p_{00.1} &= S p_{00.0} \\ p_{10.1} &= S p_{10.0} \\ p_{01.0} &= T p_{01.1} \\ p_{11.0} &= T p_{11.1} \end{aligned} \tag{5.39}$$

where S and T represent the ratios

$$\begin{aligned} S &= \frac{p_{00.1}}{p_{00.0}} = \frac{p_{10.1}}{p_{10.0}} \\ T &= \frac{p_{01.0}}{p_{01.1}} = \frac{p_{11.0}}{p_{11.1}} \end{aligned}$$

From the condition of positive effects,

$$p_{11.1} + p_{01.1} - p_{11.0} - p_{01.0} \geq 0$$

which, from Eq. (5.39), may be rewritten

$$(1 - T)(p_{11.1} + p_{01.1}) \geq 0 \tag{5.40}$$

This implies that $T \leq 1$.

Likewise, we may use the equalities in Eq. (5.39) to rewrite the probabilistic constraints given by Eqs. (5.19) and (5.20):

$$\begin{aligned} p_{00.0} + T p_{01.1} + p_{10.0} + T p_{11.1} &= 1 \\ S p_{00.0} + p_{01.1} + S p_{10.0} + p_{11.1} &= 1 \end{aligned}$$

Taking the difference of these two equations gives

$$(1 - S)(p_{00.0} + p_{10.0}) = (1 - T)(p_{01.1} + p_{11.1})$$

$T \leq 1$ then implies that $S \leq 1$.

Applying these bounds on S and T to Eq. (5.39) results in the constraints

$$\begin{aligned} p_{00.0} &\geq p_{00.1} \\ p_{10.0} &\geq p_{10.1} \\ p_{01.1} &\geq p_{01.0} \\ p_{11.1} &\geq p_{11.0} \end{aligned}$$

which, when conjoined with the conditions in Tables 5.1 and 5.2, reveal that the only applicable bounds on $\text{ACE}(D \rightarrow Y)$ under the assumption of positive effects and conditional independence are the natural bounds:

$$\begin{aligned} L_{D\rightarrow Y}(\vec{p}) &= p_{11.1} + p_{00.0} - 1 \\ &= \text{ACE}(Z \rightarrow Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\ U_{D\rightarrow Y}(\vec{p}) &= 1 - p_{01.1} - p_{10.0} \\ &= \text{ACE}(Z \rightarrow Y) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0) \end{aligned}$$

When p is not strictly positive, we can proceed through a similar exercise on a case-by-case basis and obtain identical results. We omit this part of the proof.

□

Figure 5.4 shows how the conditional independence tightens the lower bounds shown in Figure 5.3 when the only information known about the observed distribution is $\text{ACE}(Z \rightarrow D)$ and $\text{ACE}(Z \rightarrow Y)$.

5.6.2 Treatment sufficiency with structural stability

Where treatment sufficiency holds under a variety of experimental conditions, it is reasonable to assume that it is not caused by incidental equality of parameters, but rather by structural constraints. This notion of structural stability is indeed the pivotal assumption behind the causal inference methods of [PV91, SGS91], namely, that every conditional independence shown in the data must be logically implied by the decomposition of the joint probability distribution given by Eq. (5.2) as dictated by the graph structure. If this assumption holds, then the data are *DAG-isomorphic* to the graph structure, and all independence relations may then be tested by using the *d-separation* criterion ([Pea88]).

Theorem 5.6.2 *If an observed distribution $P(y, d|z)$ is structurally stable and satisfies $Y \perp\!\!\!\perp Z|D$ and $Y \not\perp\!\!\!\perp Z$, then*

$$\text{ACE}(D \rightarrow Y) = P(y_1|d_1) - P(y_1|d_0) \tag{5.41}$$

Proof:

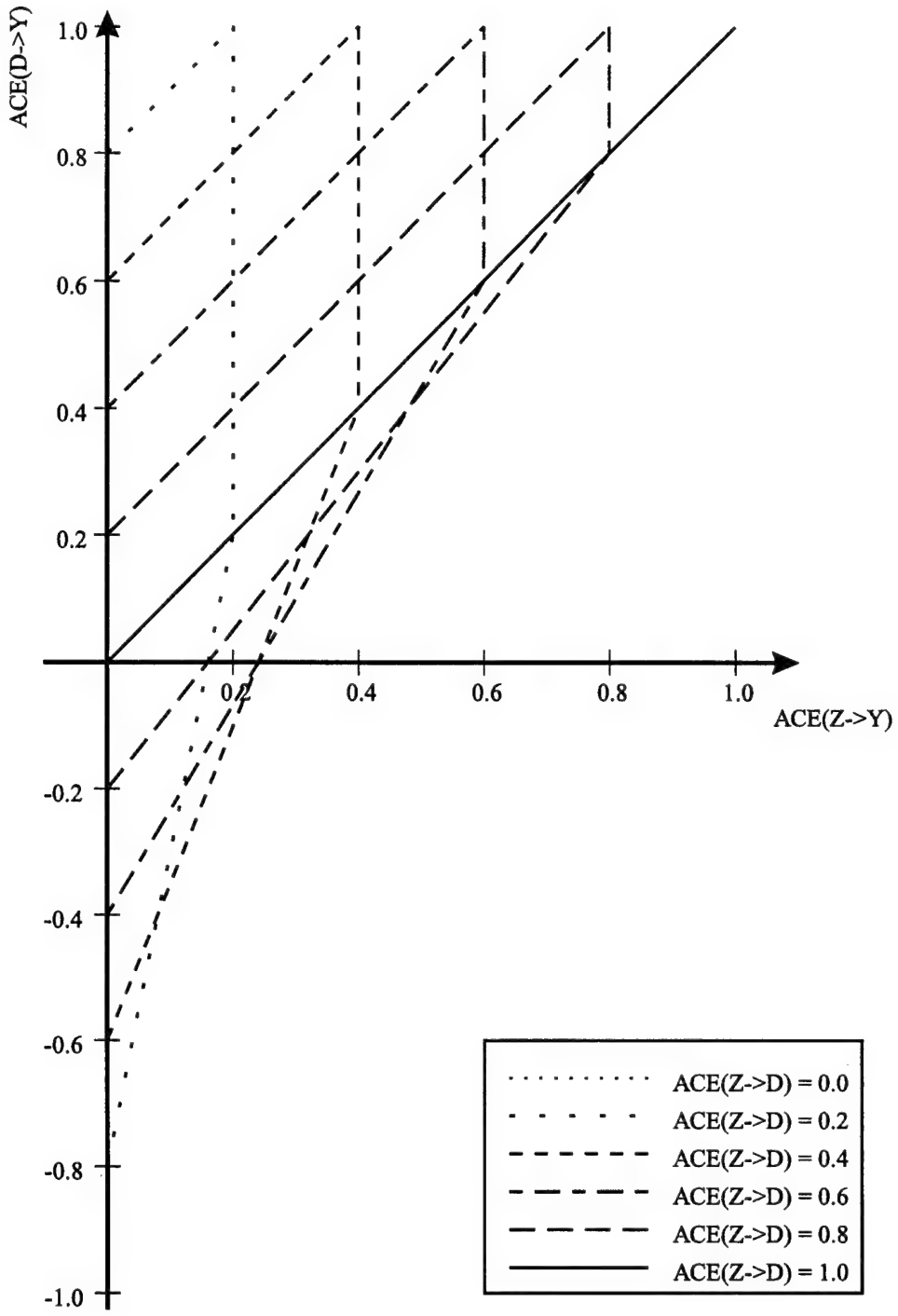

Figure 5.4: *Bounds on* $\mathrm{ACE}(D \to Y)$ *plotted against* $\mathrm{ACE}(Z \to Y)$ *and* $\mathrm{ACE}(Z \to D)$*, given that* Z *and* Y *are independent given* D*.*

The antecedent of the theorem implies that Z and Y must be d-separated given D in the graph structure for which the data is DAG-isomorphic. Applying the d-separation criterion to the graphical structure of Figure 5.1, we find that, given D, Z and Y are dependent via the path, $Z - D - U - Y$. The only way to remove this dependency is to eliminate one of the following edges: $Z \rightarrow D$, $U \rightarrow D$, or $U \rightarrow Y$. The assumption that Z and D are marginally dependent prevents the elimination of $Z \rightarrow D$; therefore, the antecedent of the theorem can only be satisfied if at least one of the edges $U \rightarrow D$ or $U \rightarrow Y$ is eliminated.

First, assume that $U \rightarrow Y$ is eliminated from the graph structure. In this case, $P(y|d,u) = P(y|d)$, which, when substituted into Eq. (5.5), results in

$$\text{ACE}(D \rightarrow Y) \;=\; P(y_1|d_1) - P(y_1|d_0)$$

Next, assume that $U \rightarrow D$ is eliminated from the graph structure. In this case, we note that $P(u) = P(u|d)$, allowing the following transformations of Eq. (5.5):

$$\begin{aligned}
\text{ACE}(D \rightarrow Y) &= \sum_u [P(u)P(y_1|d_1,u) - P(u)P(y_1|d_0,u)] \\
&= \sum_u [P(u|d_1)P(y_1|d_1,u) - P(u|d_0)P(y_1|d_0,u)] \\
&= \sum_u [P(y_1,u|d_1) - P(y_1,u|d_0)] \\
&= P(y_1|d_1) - P(y_1|d_0)
\end{aligned}$$

□

Notice that the combination of structural stability and treatment sufficiency subsumes the assumption of Eq. (5.1); $Z \perp\!\!\!\perp Y|\{D,U\}$ is no longer an assumption but is implied by $Z \perp\!\!\!\perp Y|D$, because, for any set of variables S, $Z \perp\!\!\!\perp Y|S$ cannot hold if there is a direct arc from Z to Y. Therefore, when structural stability holds, finding a variable Z' satisfying $Z' \perp\!\!\!\perp Y|D$ and $Z' \not\perp\!\!\!\perp Y$ permits us to dispose of the randomized assignment altogether and infer causal effects (using Eq. (5.41)) in purely observational studies. Discovering a Z' which satisfies these relationships may be viewed as uncovering a randomized experiment that is conducted by Nature itself, and this is the basis of the "virtual control" condition discussed in [PV91].

5.6.3 Non-defiance

A subject is characterized as *perfectly defiant* if under either treatment assignment the subject fails to comply with the assignment ($d = d_1$ if and only if $z = z_0$). In terms of the potential-response model of Figure 5.2, this behavior is specified by $r_d = 2$ in Eq. (5.14). One could imagine individuals who despise having decisions made for them. It is possible that the act of assigning them to a treatment will lead them to evade that treatment, where alternatively, they would have voluntarily selected that treatment. Consider a study that involves observation of draft status (Z) and military service (D) ([AIR93]). It is conceivable that there could be subjects who despise authority and so, if drafted, would evade service and, if not drafted, would volunteer for service.

Alternatively, there are situations in which perfectly defiant behavior would be improbable:

- when subjects do not know exactly what the two treatment options (z_0 and z_1) are; hence, it is beyond their means to defy both treatment assignments.
- when subjects know what the two treatment options are, but do not know which treatment they have been assigned (the procedures for receiving the assigned treatments are identical, as in the use of placebo).
- when subjects know what both treatments are and know which treatment they have been assigned but do not have access to both treatments; therefore, it is beyond their means to obtain the opposite treatment under either assignment.

Drug studies often are very likely to fit one of these situations, especially since a placebo is usually used as the alternative treatment to the medication under study, so subjects cannot easily determine which treatment they have been assigned.

Based on the applicability suggested above, we will define the assumption of *non-defiance* as stating that there are no perfectly defiant subjects in a study. This assumption is expressed by the constraint $P(r = r_2) = 0$, or $q_{2j} = 0$ for $j = 0, \ldots, 3$. Non-defiance together with the condition of positive effects is equivalent to the assumption of "monotonicity" analyzed by [AIR93], which translates to the restriction: either $P(r = r_2) = 0$ or $P(r = r_1) = 0$. Because the assumption of non-defiance imposes restrictions on the unobserved parameters in

Q space, it carries the potential of improving the bounds on $\text{ACE}(D \rightarrow Y)$ beyond those of Eqs. (5.27) and (5.28). The following theorem refutes this possibility.

Theorem 5.6.3 *If all subjects in a population are non-defiant, then the natural bounds on* $\text{ACE}(D \rightarrow Y)$,

$$\begin{aligned} \text{ACE}(D \rightarrow Y) &\geq \text{ACE}(Z \rightarrow Y) - P(y_1, d_0|z_1) - P(y_0, d_1|z_0) \\ \text{ACE}(D \rightarrow Y) &\leq \text{ACE}(Z \rightarrow Y) + P(y_0, d_0|z_1) + P(y_1, d_1|z_0) \end{aligned}$$

are tight.

This theorem may be proven by reapplying the linear optimization procedure detailed in Appendix B to the optimization problem given by Eq. (5.26) with the additional constraints $q_{2j} = 0$ for $j = 0, \ldots, 3$. This procedure results in a single expression each for the lower and upper bounds, identical to the natural bounds given by Eq. (5.12).

It is important to understand that the non-defiance assumption (as well as that of treatment sufficiency) does not widen the bounds of Eqs. (5.27) and (5.28) to the natural bounds, but instead restricts the observation space P to a region where the natural bounds are the only applicable bounds. Consequently, the assumption of non-defiance is partly observable; if $P(y, d|z)$ does not satisfy the following constraints implied by non-defiance

$$\begin{aligned} p_{00.0} &\geq p_{00.1} \\ p_{01.1} &\geq p_{01.0} \\ p_{10.0} &\geq p_{10.1} \\ p_{11.1} &\geq p_{11.0} \end{aligned}$$

then the assumption of non-defiance does not hold. To summarize, the assumption of non-defiance provides no benefits over the unconditional bounds given by Eqs. (5.29) and (5.30); however, it narrows the space of observation so as to render the natural bounds of Eq. (5.12) realizable.

5.6.4 Monotonic compliance and response behaviors

What if we assume that both compliance and treatment response behaviors specify monotonic functions from treatment assignment to treatment consumed and

treatment consumed to treatment response, respectively? This just corresponds to the incorporation of two additional constraints:

$$\begin{aligned} P(r_d = 2) &= 0 \\ P(r_y = 2) &= 0 \end{aligned} \tag{5.42}$$

The set of points in P space consistent with these assumptions is given by the following set of constraints:

$$\begin{aligned} p_{10.1} &\leq p_{10.0} \\ p_{01.0} &\leq p_{01.1} \\ p_{10.0} + p_{11.0} &\leq p_{10.1} + p_{11.1} \\ p_{01.0} + p_{11.0} &\leq p_{01.1} + p_{11.1} \end{aligned}$$

These last two inequalities just correspond to the positive effects convention ($\text{ACE}(Z \rightarrow Y) \geq 0$ and $\text{ACE}(Z \rightarrow D) \geq 0$).

In terms of the Q space parameters, the average treatment effect under the monotonicity assumption reduces to

$$\text{ACE}(D \rightarrow Y) = q_{01} + q_{11} + q_{31} \tag{5.43}$$

If we incorporate the constraints given by Eq. (5.42) and optimize the objective function (Eq. (5.43) we generate the following bounds on the average causal effect under the monotonicity assumption:

$$p_{10.1} + p_{11.1} - p_{10.0} - p_{11.0} \leq \text{ACE}(D \rightarrow Y) \leq 1 - p_{01.1} - p_{10.0}$$

or

$$\text{ACE}(Z \rightarrow Y) \leq \text{ACE}(D \rightarrow Y) \leq \text{ACE}(Z \rightarrow Y) + p_{11.0} + p_{00.1}$$

This shows that the average treatment effect evaluated under the monotonicity assumption is always at least as great as the causal effect evaluated from the intent-to-treat analysis.

5.7 Additional Results

5.7.1 Local average-treatment effect

While this chapter focuses primarily on predicting the average treatment effect over an entire population, there are cases where one would be interested in treatment effects averaged over a subpopulation of special characteristics. [AIR93]

have found that, under the assumption of non-defiance, the treatment effect averaged over the subpopulation of perfectly complying individuals, $\text{ACE}_c(D \to Y)$, can be identified and is given by the Instrumental Variable formula

$$\text{ACE}_c(D \to Y) = \frac{\text{ACE}(Z \to Y)}{\text{ACE}(Z \to D)} = \frac{P(y_1|z_1) - P(y_1|z_0)}{P(d_1|z_1) - P(d_1|z_0)} \tag{5.44}$$

In other words, Eq. (5.44) gives the correct treatment effect for those individuals whose participation in the treatment D comes as a consequence of the encouragement Z.

This can be verified by noting that a compliant subpopulation is characterized by the condition $r_d = 1$; thus

$$\begin{aligned}
\text{ACE}_c(D \to Y) &= P(y_1|d_1, r_d{=}1) - P(y_1|d_0, r_d{=}1) \\
&= P(r_y{=}1|r_d{=}1) - P(r_y{=}2|r_d{=}1) \\
&= \frac{P(r_d{=}1, r_y{=}1) - P(r_d{=}1, r_y{=}2)}{P(r_d{=}1)} \\
&= \frac{q_{11} - q_{12}}{q_{10} + q_{11} + q_{12} + q_{13}}
\end{aligned}$$

This last expression coincides with the Instrumental Variable formula above under the condition of non-defiance, namely, $P(r_d{=}2) = 0$, or $q_{2j} = 0$ for $j = 0, \ldots, 3$.

It is worth noting that the subpopulation of perfectly complying individuals is not, in general, identifiable, because the condition $r_d = 1$ cannot be determined from the triplet (y, d, z). Nevertheless, the behavior of this subpopulation may be of interest to analysts, as it reveals the treatment effect under ideal conditions, free of noncompliance side effects. Bounds on the behavior of other subpopulations of interest can be obtained by methods similar to those in Section 5.2.2.

5.7.2 Treatment effect given treatment consumed

Some researchers might claim that the average treatment effect ($\text{ACE}(D \to Y)$) is not the deciding factor when developing a policy for patient care, because they believe that patients' compliance behaviors in clinical practice will be similar to their compliance during drug trials. Although the author disagrees with this interpretation — patients would be more apt to follow the advice of their physician in taking an approved drug demonstrated to be effective — this section will examine the conditional treatment effect that would be the basis for their policy decision.

If patients act in clinical practice as they do in drug studies, then the researcher is actually interested in the average causal effect of drug treatment for those subjects who actually consumed the drug (d_1). In other words, derive bounds for $P(y_1^*|\hat{d}_1^*, d_1) - P(y_1^*|\hat{d}_0^*, d_1)$ given the observed distribution $P(z, d, y)$.

These counterfactual probabilities may be written in terms of the distribution of respons functions:

$$P(y_1^*|\hat{d}_0^*, d_1) = \frac{P(z_1)[q_{12} + q_{13} + q_{32} + q_{33}] + P(z_0)[q_{22} + q_{23} + q_{32} + q_{33}]}{P(d_1)}$$

$$P(y_1^*|\hat{d}_1^*, d_1) = \frac{P(z_1)[q_{11} + q_{13} + q_{31} + q_{33}] + P(z_0)[q_{21} + q_{23} + q_{31} + q_{33}]}{P(d_1)}$$

Taking the difference between these two expressions gives us the average causal effect of treatment on response for those individuals who took the treatment:

$$\text{ACE}(D \rightarrow Y|d_1) = \frac{P(z_1)[q_{11} + q_{31} - q_{12} - q_{32}] + P(z_0)[q_{21} + q_{31} - q_{22} - q_{32}]}{P(d_1)} \tag{5.45}$$

Given a specific distribution for the treatment assignment $P(z)$, we can apply linear symbolic optimization to the numerator of this equation. It turns out that the Q space expressions multiplied by $P(z_1)$ and $P(z_0)$ may be optimized independently. This can be shown by deriving the closed-form bounds on $f_1(\vec{q}) = q_{11} + q_{31} - q_{12} - q_{32}$ and $f_2(\vec{q}) = q_{21} + q_{31} - q_{22} - q_{32}$ and demonstrating that the sum of their lower (upper) bounds is equal to the closed-form lower (upper) bound on $f_1(\vec{q}) + f_2(\vec{q})$. Since the coefficients on f_1 and f_2 ($P(z_1)$ and $P(z_0)$, respectively) are non-negative, we can optimize Eq. (5.45) by optimizing f_1 and f_2 independently.

Following this strategy, closed-form bounds on the conditional average causal effect may be derived, and are expressed in terms of the distribution of observables $P(y, d, z)$:

$$\frac{1}{P(d_1)} \max \left\{ \begin{array}{c} P(z_0)[p_{10.0} + p_{11.0} - p_{10.1} - p_{11.1}] - p_{01.1} \\ P(z_1)[p_{10.1} + p_{11.1} - p_{10.0} - p_{11.0}] - p_{01.0} \\ P(z_0)[p_{11.0} - p_{10.1} - p_{11.1}] - P(z_1)p_{10.0} - p_{01.0} \\ P(z_1)[p_{11.1} - p_{10.0} - p_{11.0}] - P(z_0)p_{10.1} - p_{01.1} \end{array} \right\}$$

$$\leq \text{ACE}(D \rightarrow Y|d_1) \leq$$

$$\frac{1}{P(d_1)} \min \left\{ \begin{array}{c} P(z_0)p_{11.0} + P(z_1)[p_{10.1} + p_{11.1} - p_{10.0}] \\ P(z_0)[p_{10.0} + p_{11.0} - p_{10.1}] + P(z_1)p_{11.1} \\ P(z_1)[p_{00.0} + p_{01.0} - p_{01.1}] + P(z_0)p_{00.1} + p_{11.1} \\ P(z_1)p_{00.0} + p_{11.0} + P(z_0)[p_{00.1} + p_{01.1} - p_{01.0}] \end{array} \right\}$$

or

$$\frac{1}{P(d_1)} \max \left\{ \begin{array}{l} P(y_1, z_0) - P(y_1|z_1)P(z_0) - P(y_0, d_1|z_1) \\ P(y_1, z_1) - P(y_1|z_0)P(z_1) - P(y_0, d_1|z_0) \\ P(y_1, z_0) - P(y_1|z_1)P(z_0) - P(y_1, d_0|z_0) - P(y_0, d_1|z_0) \\ P(y_1, z_1) - P(y_1|z_0)P(z_1) - P(y_1, d_0|z_1) - P(y_0, d_1|z_1) \end{array} \right\}$$

$$\leq \text{ACE}(D \rightarrow Y|d_1) \leq$$

$$\frac{1}{P(d_1)} \min \left\{ \begin{array}{l} P(y_1, z_1) - P(y_1|z_0)P(z_1) + P(y_1, d_1|z_0) \\ P(y_1, z_0) - P(y_1|z_1)P(z_0) + P(y_1, d_1|z_1) \\ P(y_1, z_1) - P(y_1|z_0)P(z_1) + P(y_0, d_0|z_1) + P(y_1, d_1|z_1) \\ P(y_1, z_0) - P(y_1|z_1)P(z_0) + P(y_0, d_0|z_0) + P(y_1, d_1|z_0) \end{array} \right\}$$

5.7.3 Divergence of intent-to-treat analysis from treatment effect bounds

Strides have been made to educate the scientific community to the potential errors in evaluating treatment effects from the intent-to-treat analysis, $\text{ACE}(Z \rightarrow Y) = P(y_1|z_1) - P(y_1|z_0)$; however, there are still some who incorrectly use this expression to evaluate a drug's efficacy in a quasi-experimental study. This begs the question, just how inaccurate is the intent-to-treat analysis? Even though this analysis does not produce the correct bounds for the treatment effect, does the computed value at least provide a feasible value for the treatment effect? Unfortunately, not always. In fact, this divergence from the bounds can occur even in cases where $\text{ACE}(Z \rightarrow D)$ approaches 100%. For example, consider the point in P space:

$$\begin{array}{ll} p_{00.0} = 0.84 & p_{00.1} = 0.08 \\ p_{01.0} = 0.16 & p_{01.1} = 0.00 \\ p_{10.0} = 0.00 & p_{10.1} = 0.12 \\ p_{11.0} = 0.00 & p_{11.1} = 0.80 \end{array}$$

The compliance is given by the average causal effect of treatment assignment on treatment received

$$\text{ACE}(Z \rightarrow D) \;=\; 0.64$$

and the bounds on the average causal effect are computed to be

$$0.68 \leq \text{ACE}(D \rightarrow Y) \leq 0.72$$

Hoever, the intent-to-treat analysis gives

$$\text{ACE}(Z \rightarrow Y) \;=\; 0.92$$

Here we see that even though compliance is relatively high, the intent-to-treat analysis can lead to results outside the bounds on the average causal effect.

In general, $\text{ACE}(Z \rightarrow Y)$ will fall below $\text{ACE}(D \rightarrow Y)$'s lower bound in the following three regions of P space:

$$\begin{array}{|rcl|}
\hline
p_{01.0} + p_{10.0} & \geq & p_{01.1} \\
p_{01.1} + p_{10.1} & \geq & p_{10.0} \\
p_{00.1} + p_{11.0} & \geq & p_{10.1} + p_{11.1} + p_{00.0} + p_{01.0} \\
\hline
p_{01.1} & \geq & p_{01.0} + p_{10.0} \\
p_{11.0} & \geq & p_{10.1} + p_{11.1} + \frac{1}{2}p_{01.0} \\
\hline
p_{10.0} & \geq & p_{01.1} + p_{10.1} \\
p_{00.1} & \geq & p_{00.0} + p_{01.0} + \frac{1}{2}p_{10.1} \\
\hline
\end{array}$$

In addition, $\text{ACE}(Z \rightarrow Y)$ rises above $\text{ACE}(D \rightarrow Y)$'s upper bound in the following three regions of P space:

$$\begin{array}{|rcl|}
\hline
p_{00.0} + p_{11.0} & \geq & p_{11.1} \\
p_{00.1} + p_{11.1} & \geq & p_{00.0} \\
p_{01.0} + p_{10.1} & \geq & p_{10.0} + p_{11.0} + p_{00.1} + p_{01.1} \\
\hline
p_{11.1} & \geq & p_{00.0} + p_{11.0} \\
p_{01.0} & \geq & p_{00.1} + p_{01.1} + \frac{1}{2}p_{11.0} \\
\hline
p_{00.0} & \geq & p_{00.1} + p_{11.1} \\
p_{10.1} & \geq & p_{10.0} + p_{11.0} + \frac{1}{2}p_{00.1} \\
\hline
\end{array}$$

In all other regions of P space consistent with the canonical partial compliance model, ACE($Z \rightarrow Y$) will fall within ACE($D \rightarrow Y$)'s upper and lower bounds.

Therefore, the intent-to-treat analysis not only fails to reflect the uncertainty imposed by subject noncompliance, but may also lead to an estimate of the causal effect that lies outside its actual bounds. In other words, the intent-to-treat analysis can not even state that its estimate is potentially correct given the observed distribution.

5.8 Conclusions

This chapter provided formulas that allow analysts to make categorical statements about causal effects in the context of studies where subjects are only partially compliant. These formulas, expressed in terms of the distribution over observed variables (treatment assignment, treatment received, and observed response), represent strict upper and lower bounds for the average causal effect of the treatment on the population. These bounds are applicable to all studies where the assignment itself only affects the observed response via the treatment actually received, regardless of any interaction that might take place between the treatment received and the observed response. Aside from this assumption, the results do not rest on any particular model of compliance behavior.

We believe that the results presented here could be particularly helpful in quasi-experimental studies, that is, studies in which randomized mandated treatments are either unfeasible or undesirable and randomized encouragements are instituted instead ([Hol88]). For example, in evaluating the efficacy of a social program, the randomized instrument can be advertisement, incentives, or eligibility, letting subjects make the final choice of participation. The bounds established through Eqs. (5.29) and (5.30) reveal that such studies, despite the indirectness of the randomized instrument, can yield valuable information on the average causal effect of the treatment on the population.

One topic that should receive attention in future work is the maximum-likelihood estimation technique for finite samples.

CHAPTER 6

Continuous treatments

6.1 Introduction

In the last chapter, strict upper and lower bounds on the causal effect of treatment on response from partial compliance studies were derived using linear optimization techniques. These bounds were derived for a model where the set of observed variables (treatment assignment, treatment received, and observed response) are all binary. Aside from the qualitative structure of the model, those results are assumption free. [BP93, Sections 1 and 2] and [Pea93a] provide motivation for studying the causal effects identification task and explain the basic qualitative assumptions which are applied in this chapter to derive results applicable when the received treatment variable is not binary.

When the observed received treatment is not binary, it is difficult, if not impossible, to translate the causal effect bounds evaluated from a binary model into a policy statement. For example, consider a quasi-experiment where subjects are encouraged to take either two units of treatment or zero units of treatment. At the end of data collection we find that besides zero and two units, many subjects consumed just one unit of treatment. In order to apply the bounds of Section 5.3 the received treatment must be transformed to a binary variable. In one attempt, the one and two unit treatments are merged into the positive received treatment category and the resulting distribution is substituted into Eqs. (5.29) and (5.30) to compute the average treatment effect. After this analysis we might find that the lower bound on the treatment causal effect is positive; therefore, a strict treatment policy is developed which states that patients who meet the studies selection criterion will be forced to consume two units of treatment.

Unfortunately, this transformation of the three value domain to a two value domain loses information about the distribution of received treatment, in particular, between one and two units. It is possible that relatively few subjects consumed two units of treatment, and for those subjects the treatment had a negative causal effect on response. At the same time, the treatment causal effect

was strong for those subjects who consumed just one unit of treatment. Therefore, if a treatment policy is implemented that forces two unit consumption, then subjects will suffer negative consequences on average. Because of this shortcoming of the binary treatment analysis, this chapter will further partition the received treatment domain such that meaningful and safe results may be obtained for continuous received treatment data.

[AI92] and [EF91] have looked at the analysis of causal effects for studies where the domain of the received treatment variable is not binary. In [EF91] the partial compliance data is fit by a naive treatment-response curve for both the placebo and treatment. The actual treatment-response curve is then related to these two measurable curves along with other unmeasurable factors. Specific assumptions allow estimation of the actual treatment-response curve, but in general, this is not possible. Their framework differs from the model presented here, in that the observed response variable in [EF91] is continuous allowing specification of a treatment-response curve, while our use of a binary observed response variable only allows us to specify the probability that the response will fall within a particular range. [AI92] partition the continuous treatment variable and show that under a condition of monotonicity the treatment causal effect can be determined for the class of subjects whose treatment is influenced by their treatment assignment. Their partitioning of the treatment variable and direct use of the continuous treatment response allows evaluation of a treatment-response curve.

Section 6.2 describes the received treatment partitioning strategy which enables derivation of bounds on the causal effect of one treatment level versus a base treatment level when the domain of received treatment is continuous. An example demonstrating the application of those closed-form bounds will then be presented in Section 6.3. In Section 6.4 we will demonstrate that these bounds may be further tightened when more than two homogeneous treatment levels are extracted from the continuous domain. Section 6.5 presents some concluding remarks.

6.2 Derivation of continuous treatment bounds

Suppose that the domain of the treatment variable D is no longer binary, but is now continuous. We are interested in the average causal effect of one level of treatment d_0 (the control) versus the effect of another level d_1 (nominal treatment). All other treatments in D's domain not in $\{d_0, d_1\}$ will be labelled by d_m.

Because d_0 and d_1 coincide with exact values of treatment, the independence relations discussed in Chapter 5 still hold:

$$Z \perp\!\!\!\perp Y \mid \{D = d_0, U\}$$
$$Z \perp\!\!\!\perp Y \mid \{D = d_1, U\}$$

However, Z and Y are no longer independent given U and $D = d_m$:

$$Z \not\perp\!\!\!\perp Y \mid \{D = d_m, U\}$$

We can derive bounds on the average causal effect $\text{ACE}(D \rightarrow Y)$ by first specifying the model not in terms of a completely functional model, but in terms of a partial functional model. Because D is continuous, we cannot completely specify the response function r_y mapping D to Y. Instead we specify the *partial response function* mapping only part of D's domain (d_0 and d_1) to Y:

$$y = f_y(d, r_y) = h_{y,r_y}(d)$$

where

$$h_{y,0}(d) = \begin{cases} y_0 & \text{if } d \in \{d_0, d_1\} \\ \text{undef} & \text{if } d = d_m \end{cases}$$

$$h_{y,1}(d) = \begin{cases} y_0 & \text{if } d = d_0 \\ y_1 & \text{if } d = d_1 \\ \text{undef} & \text{if } d = d_m \end{cases}$$

$$h_{y,2}(d) = \begin{cases} y_1 & \text{if } d = d_0 \\ y_0 & \text{if } d = d_1 \\ \text{undef} & \text{if } d = d_m \end{cases}$$

$$h_{y,3}(d) = \begin{cases} y_1 & \text{if } d \in \{d_0, d_1\} \\ \text{undef} & \text{if } d = d_m \end{cases}$$

D is still functionally defined by

$$d = f_d(z, r_d) = h_{d,r_d}(z) \tag{6.1}$$

where

$$h_{d,0}(z) = d_0$$
$$h_{d,1}(z) = \begin{cases} d_0 & \text{if } z = z_0 \\ d_m & \text{if } z = z_1 \end{cases}$$

$$
\begin{aligned}
h_{d,2}(z) &= \begin{cases} d_0 & \text{if } z = z_0 \\ d_1 & \text{if } z = z_1 \end{cases} \\
h_{d,3}(z) &= \begin{cases} d_m & \text{if } z = z_0 \\ d_0 & \text{if } z = z_1 \end{cases} \\
h_{d,4}(z) &= d_m \\
h_{d,5}(z) &= \begin{cases} d_m & \text{if } z = z_0 \\ d_1 & \text{if } z = z_1 \end{cases} \\
h_{d,6}(z) &= \begin{cases} d_1 & \text{if } z = z_0 \\ d_0 & \text{if } z = z_1 \end{cases} \\
h_{d,7}(z) &= \begin{cases} d_1 & \text{if } z = z_0 \\ d_m & \text{if } z = z_1 \end{cases} \\
h_{d,8}(z) &= d_1
\end{aligned}
$$

Let $q_{jk} = P(r_d{=}j, r_y{=}k)$. Then we may write the linear relationship between the P space and the Q space as follows:

$$
\begin{aligned}
P(y_0, d_0|z_0) &= q_{00} + q_{01} + q_{10} + q_{11} + q_{20} + q_{21} \\
P(y_0, d_1|z_0) &= q_{60} + q_{62} + q_{70} + q_{72} + q_{80} + q_{82} \\
P(y_1, d_0|z_0) &= q_{02} + q_{03} + q_{12} + q_{13} + q_{22} + q_{23} \\
P(y_1, d_1|z_0) &= q_{61} + q_{63} + q_{71} + q_{73} + q_{81} + q_{83} \\
P(d_m|z_0) &= q_{30} + q_{31} + q_{32} + q_{33} + q_{40} + q_{41} + q_{42} + q_{43} + \\
&\quad q_{50} + q_{51} + q_{52} + q_{53}
\end{aligned}
$$

$$
\begin{aligned}
P(y_0, d_0|z_1) &= q_{00} + q_{01} + q_{30} + q_{31} + q_{60} + q_{61} \\
P(y_0, d_1|z_1) &= q_{20} + q_{22} + q_{50} + q_{52} + q_{80} + q_{82} \\
P(y_1, d_0|z_1) &= q_{02} + q_{03} + q_{32} + q_{33} + q_{62} + q_{63} \\
P(y_1, d_1|z_1) &= q_{21} + q_{23} + q_{51} + q_{53} + q_{81} + q_{83} \\
P(d_m|z_1) &= q_{10} + q_{11} + q_{12} + q_{13} + q_{40} + q_{41} + q_{42} + q_{43} + \\
&\quad q_{70} + q_{71} + q_{72} + q_{73}
\end{aligned}
$$

The reason why $P(y_0, d_m|z) + P(y_1, d_m|z)$ is treated as a single value $P(d_m|z)$, is that the individual components cannot be expressed in terms of the Q space parameters.

In terms of the Q parameter space we can write the average causal effect as

$$\text{ACE}(D \rightarrow Y) = \sum_{j=0}^{3} q_{j1} - q_{j2}$$

Given this objective function and linear constraints on the Q space, we may derive general upper and lower bounds on $\text{ACE}(D \rightarrow Y)$:

$$L_{D\rightarrow Y}(\vec{p}) = \max \left\{ \begin{array}{c} p_{00.0} + p_{11.1} - 1 \\ p_{00.1} + p_{11.1} - 1 \\ p_{11.0} + p_{00.1} - 1 \\ p_{00.0} + p_{11.0} - 1 \\ 2p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} - 2 \\ p_{00.0} + 2p_{11.0} + p_{00.1} + p_{01.1} - 2 \\ p_{10.0} + p_{11.0} + 2p_{00.1} + p_{11.1} - 2 \\ p_{00.0} + p_{01.0} + p_{00.1} + 2p_{11.1} - 2 \end{array} \right\} \tag{6.2}$$

$$U_{D\rightarrow Y}(\vec{p}) = \min \left\{ \begin{array}{c} 1 - p_{10.0} - p_{01.1} \\ 1 - p_{01.0} - p_{10.1} \\ 1 - p_{01.0} - p_{10.0} \\ 1 - p_{01.1} - p_{10.1} \\ 2 - 2p_{01.0} - p_{10.0} - p_{10.1} - p_{11.1} \\ 2 - p_{01.0} - 2p_{10.0} - p_{00.1} - p_{01.1} \\ 2 - p_{10.0} - p_{11.0} - 2p_{01.1} - p_{10.1} \\ 2 - p_{00.0} - p_{01.0} - p_{01.1} - 2p_{10.1} \end{array} \right\} \tag{6.3}$$

It is very important to understand that no assumptions whatsoever have been made about the range of d_m, or the functional mapping from any values in d_m to Y. Therefore, these bounds hold true (they might be loose if we obtain more specific information about d_m) regardless of the composition of d_m.

Often, in the real world, practically no subjects will consume exactly d_0 or d_1 units of treatment. Therefore, we must make an assumption that there exists homogeneous treatment windows around d_0 and d_1. In other words, if any subject forced to consume d_1 (d_0) units of treatment has a response of $Y = y$, then if the subject would have consumed an amount of treatment in $[d_1 - \delta, d_1 + \delta]$ ($[d_0 - \epsilon, d_0 - \epsilon]$), the subject would have had the same response $Y = y$. This is a reasonable assumption when the window sizes (2δ and 2ϵ) are much smaller than the difference between d_1 and d_0. If d_0 is defined as zero units of treatment, then it is desirable that the window be of zero width ($\epsilon = 0$). These will be the assumptions made in our reanalysis of the Lipid study data.

6.3 Example

We will now show by example how the bounds of Eqs. (6.2) and (6.3) can be used to provide meaningful information about causal effects. Reconsider the Lipid Research Clinics Coronary Primary Prevention Trial data described in Section 5.4.

In order to apply our analysis to this study, the continuous data obtained in the [Pro84] study must be transformed to the discrete variables representing *treatment assignment* (Z), *received treatment* (D), and *observed response* (Y). The following transformation accomplishes this by thresholding dosage consumption and change in cholesterol level:

$$d = \begin{cases} d_0 & \text{if } z = z_0 \text{ or } b = 0 \\ d_1 & \text{if } z = z_1 \text{ and } \gamma - \rho \leq b \leq \gamma + \rho \\ d_m & \text{otherwise} \end{cases} \tag{6.4}$$

$$y = \begin{cases} y_0 & \text{if } c_I - c_F < \delta \\ y_1 & \text{if } c_I - c_F \geq \delta \end{cases} \tag{6.5}$$

γ and ρ are the center and radius of the window of positive treatment, while δ specifies the minimum decrease in cholesterol level which we consider a positive treatment. This discretization assumes that subjects taking between $\gamma - \rho$ and $\gamma + \rho$ units of cholestyramine form a homogeneous treatment-response group. In addition, Eq. (6.4) reflects the finding that subjects assigned placebo (z_0) did not take cholestyramine, namely,

$$\begin{aligned} P(d_1|z_0) &= 0 \\ P(d_m|z_0) &= 0 \end{aligned}$$

Clearly, by varying this threshold over the range of Y one obtains upper and lower bounds on the entire distribution of the treatment effect, $P(Y^* \leq y|\hat{d}_1) - P(Y^* \leq y|\hat{d}_0)$.

For the current analysis we set $\rho = 7$ and $\gamma = 94$, while the threshold for cholesterol level reduction in Eq. (6.5) was selected at $\delta = 38$ units. If the data samples are interpreted according to (6.4) and (6.5), then the conditional distribution over (Z, D, Y) results in the distribution given in Table 6.1[1]

By computing the quantities required for (6.2), we obtain

$$\text{ACE}(D \rightarrow Y) \geq \max \left\{ \begin{array}{l} 0.262, -0.685, -0.976, -0.029, \\ 0.233, -0.902, -1.632, -0.423 \end{array} \right\} = 0.262$$

[1]We make the large-sample assumption and take the sample frequencies as representing $P(y, d|z)$.

$P(y,d\|z)$	z_0		z_1	
	y_0	y_1	y_0	y_1
d_0	0.971	0.029	0.024	0.000
d_m	0.000	0.000	0.436	0.146
d_1	0.000	0.000	0.103	0.291

Table 6.1: Conditional probability distribution $P(y,d|z)$ for the Lipid Research Clinic Program (1984) data, discretized by Eqs. (6.4) and (6.5).

Those needed for (6.3) give us

$$\text{ACE}(D \rightarrow Y) \leq \min \left\{ \begin{array}{l} 0.868, 1.000, 0.971, 0.897, \\ 1.680, 1.815, 1.765, 0.926 \end{array} \right\} = 0.868$$

Accordingly, we conclude that the treatment causal effect lies in the range

$$0.262 \leq \text{ACE}(D \rightarrow Y) \leq 0.868$$

which is quite informative; the experimenter can categorically state that when applied uniformly to the population, a dosage of 84 to 101 units of cholestyramine is guaranteed to improve by at least 26.2% the probability of reducing a patient's level of cholesterol by 38 points or more. This guarantee is established despite the fact that 60.6% of the subjects in the treatment group did not comply with their assigned dosage level. For comparison, note that the intent-to-treat analysis in this study gives $P(y_1|z_1) - P(y_1|z_0) = 0.408$, meaning that enforcing full compliance might result in as much as 26% improvement and no more than 14.6% reduction in the proportion of patients benefiting from the treatment.

In the above analysis, we selected ρ and γ such that the subjects who consumed the greatest quantity of cholestyramine would be classified as having received positive treatment. This is not necessary, though; beyond a certain dosage, a treatment may actually impede the mechanism whereby positive response is attained. It is possible that a higher feasible range of causal effects may be attainable by examining a different range of consumed treatment than the maximum range. Hence, we can re-analyze the cholesterol treatment by evaluating the feasible range of $\text{ACE}(D \rightarrow Y)$ for different values of γ while keeping ρ and δ fixed. Figure 6.1 presents the results of this analysis, and shows that the highest lower bound on the treatment causal effect is obtained when we use the maximum received treatment. In a sense, this graph can be viewed as a treatment-response curve, where the differences in the probability of reducing the cholesterol level

by at least δ units under treatment and placebo are plotted against the received treatment (γ), rather than a plot of the difference in cholesterol plotted against received treatment.

It is interesting to see how the bounds on ACE($D \rightarrow Y$) in this study are dependent on the threshold (δ) used to transform the continuous observed response to the binary observed response. The results in Figure 6.1 indicate that the maximum received treatment in the cholestyramine study gives the highest lower bound on the treatment causal effect (which is preferred in this case); therefore, we plot the treatment causal effect of the maximum cholestyramine dosage as a function of δ. These results are rendered in Figure 6.2.

6.4 Further decomposition of treatment

It is possible that the bounds calculated from Eqs. (6.2) and (6.3) may become tighter as the remaining treatment class d_m is further decomposed and incorporated into the analysis. For example, suppose that the treatment variable's domain may be partitioned into three variables, such that the independence assumptions are sufficiently accurate:

$$\begin{aligned} Z \perp\!\!\!\perp Y \mid \{D = d_0, U\} \\ Z \perp\!\!\!\perp Y \mid \{D = d_1, U\} \\ Z \perp\!\!\!\perp Y \mid \{D = d_m, U\} \end{aligned}$$

In this case, the domain of the treatment-response variable, r_y, may be partitioned into eight values, where Y is functionally determined by D and r_y

$$y = f_y(d, r_y) = h_{y,r_y}(d) \tag{6.6}$$

where

$$\begin{aligned} h_{y,0}(d) &= y_0 \\ h_{y,1}(d) &= \begin{cases} y_0 & \text{if } d \in \{d_0, d_m\} \\ y_1 & \text{if } d = d_1 \end{cases} \\ h_{y,2}(d) &= \begin{cases} y_0 & \text{if } d = d_0 \\ y_1 & \text{if } d = d_m \\ y_0 & \text{if } d = d_1 \end{cases} \\ h_{y,3}(d) &= \begin{cases} y_0 & \text{if } d = d_0 \\ y_1 & \text{if } d \in \{d_m, d_1\} \end{cases} \end{aligned}$$

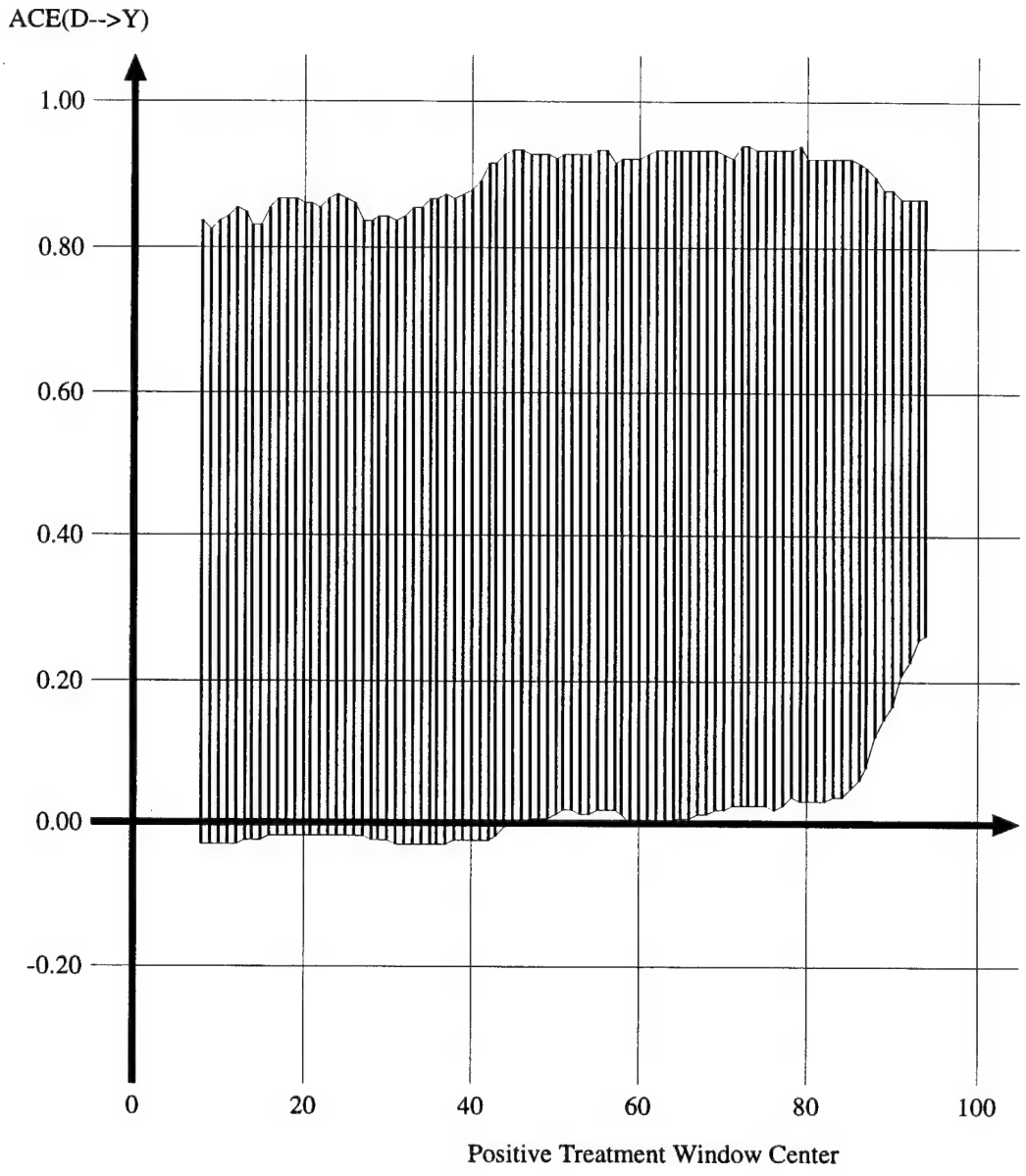

Figure 6.1: *Ranges of* ACE($D \rightarrow Y$) *evaluated for the cholestyramine treatment data for different positive treatment window centers* (γ). *For all values of* γ, *the radius of the positive treatment window* (ρ) *is 7 and the positive observed response threshold* (δ) *is 38.*

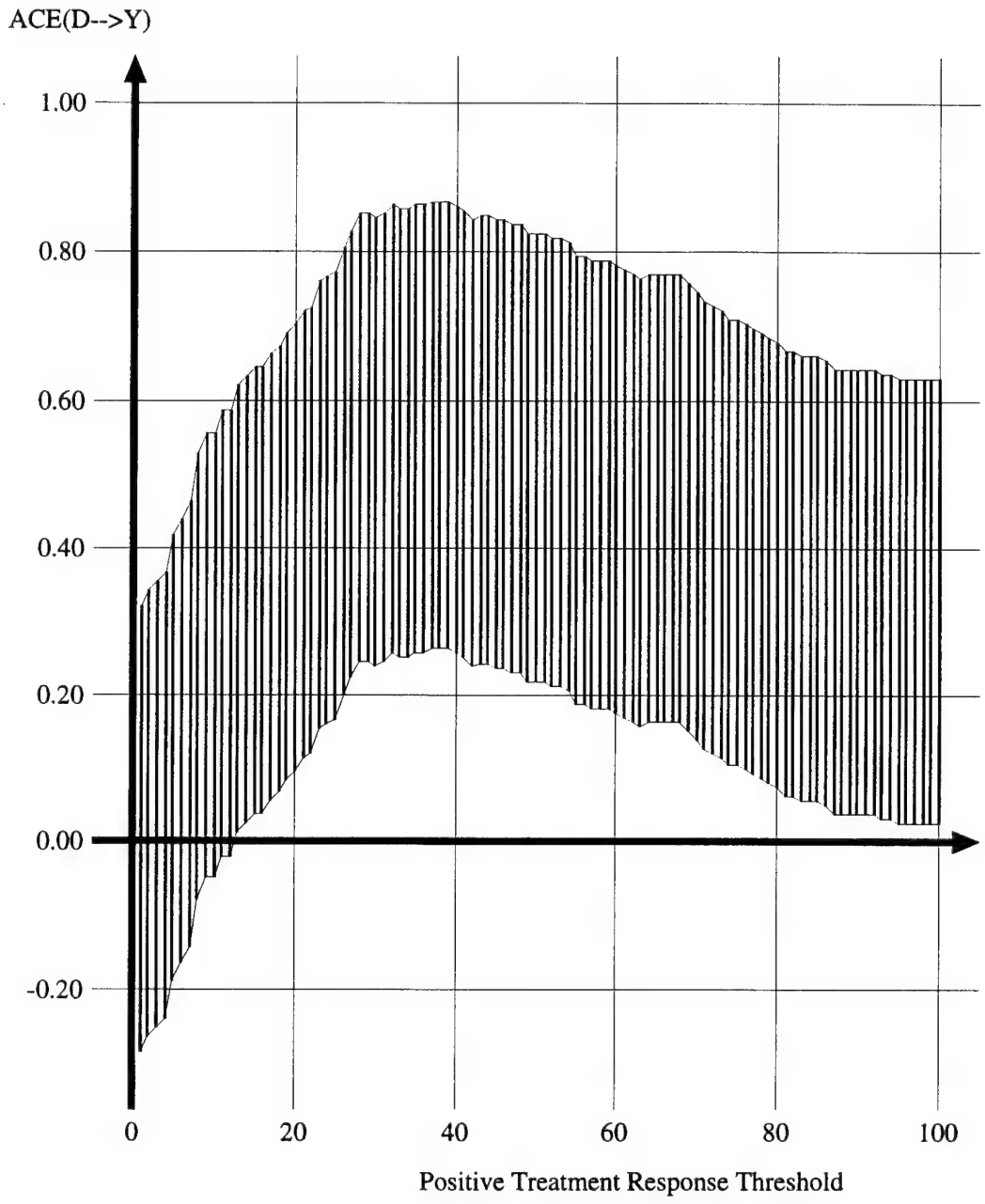

Figure 6.2: *Ranges of* ACE($D \rightarrow Y$) *evaluated for the cholestyramine treatment data for different positive observed response thresholds* (δ). *For all values of* δ, *the radius of the positive treatment window* (ρ) *is 7 and the positive treatment window center* (γ) *is 94.*

$$
\begin{aligned}
h_{y,4}(d) &= \begin{cases} y_1 & \text{if } d = d_0 \\ y_0 & \text{if } d \in \{d_m, d_1\} \end{cases} \\
h_{y,5}(d) &= \begin{cases} y_1 & \text{if } d = d_0 \\ y_0 & \text{if } d = d_m \\ y_1 & \text{if } d = d_1 \end{cases} \\
h_{y,6}(d) &= \begin{cases} y_1 & \text{if } d \in \{d_0, d_m\} \\ y_0 & \text{if } d = d_1 \end{cases} \\
h_{y,7}(d) &= y_1
\end{aligned}
$$

The causal effect of the treatment can now be obtained directly from Eqs. (6.1) and (6.6), giving

$$
\begin{aligned}
P(y_1^*|\hat{d}_1^*) &= P(r_y{=}1) + P(r_y{=}3) + P(r_y{=}5) + P(r_y{=}7) \\
P(y_1^*|\hat{d}_0^*) &= P(r_y{=}4) + P(r_y{=}5) + P(r_y{=}6) + P(r_y{=}7)
\end{aligned}
$$

and

$$
\text{ACE}(D \rightarrow Y) = P(r_y{=}1) + P(r_y{=}3) - P(r_y{=}4) - P(r_y{=}6) \qquad (6.7)
$$

The distribution over potential responses, $P(r_d, r_y)$, is specified by 72 parameters. Let these parameters be notated as follows:

$$
q_{ij} = P(r_d{=}i, r_y{=}j)
$$

The probabilistic constraint

$$
\sum_{i=0}^{8} \sum_{j=0}^{7} q_{jk} = 1
$$

implies that there are only 71 independent parameters.

In terms of the Q parameter space we can rewrite Eq. (6.7) as

$$
\text{ACE}(D \rightarrow Y) = \sum_{j=0}^{8} [q_{j1} + q_{j3} - q_{j4} - q_{j6}]
$$

The conditional distribution $P(y, d|z)$ over the observable variables is fully specified by 12 parameters, which will be notated as follows:

$$
p_{00.0} = P(y_0, d_0|z_0)
$$

$$
\begin{aligned}
p_{0m.0} &= P(y_0, d_m|z_0) \\
p_{01.0} &= P(y_0, d_1|z_0) \\
p_{10.0} &= P(y_1, d_0|z_0) \\
p_{1m.0} &= P(y_1, d_m|z_0) \\
p_{11.0} &= P(y_1, d_1|z_0) \\
p_{00.1} &= P(y_0, d_0|z_1) \\
p_{0m.1} &= P(y_0, d_m|z_1) \\
p_{01.1} &= P(y_0, d_1|z_1) \\
p_{10.1} &= P(y_1, d_0|z_1) \\
p_{1m.1} &= P(y_1, d_m|z_1) \\
p_{11.1} &= P(y_1, d_1|z_1)
\end{aligned}
$$

The probabilistic constraints

$$
\begin{aligned}
\sum_{i\in\{0,1\}} \sum_{j\in\{0,m,1\}} p_{ij.0} &= 1 \\
\sum_{i\in\{0,1\}} \sum_{j\in\{0,m,1\}} p_{ij.1} &= 1
\end{aligned}
$$

implies that there are only 10 independent parameters.

Given some point $\vec{q}$ in Q space, there is a direct linear transformation to the corresponding point $\vec{p}$ in the observation space P:

$$
\begin{aligned}
p_{00.0} &= q_{00} + q_{01} + q_{02} + q_{03} + q_{10} + q_{11} + q_{12} + q_{13} + q_{20} + q_{21} + q_{22} + q_{23} \\
p_{0m.0} &= q_{30} + q_{31} + q_{34} + q_{35} + q_{40} + q_{41} + q_{44} + q_{45} + q_{50} + q_{51} + q_{54} + q_{55} \\
p_{01.0} &= q_{60} + q_{62} + q_{64} + q_{66} + q_{70} + q_{72} + q_{74} + q_{76} + q_{80} + q_{82} + q_{84} + q_{86}
\end{aligned}
$$

$$
\begin{aligned}
p_{10.0} &= q_{04} + q_{05} + q_{06} + q_{07} + q_{14} + q_{15} + q_{16} + q_{17} + q_{24} + q_{25} + q_{26} + q_{27} \\
p_{1m.0} &= q_{32} + q_{33} + q_{36} + q_{37} + q_{42} + q_{43} + q_{46} + q_{47} + q_{52} + q_{53} + q_{56} + q_{57} \\
p_{11.0} &= q_{61} + q_{63} + q_{65} + q_{67} + q_{71} + q_{73} + q_{75} + q_{77} + q_{81} + q_{83} + q_{85} + q_{87}
\end{aligned}
$$

$$
\begin{aligned}
p_{00.1} &= q_{00} + q_{01} + q_{02} + q_{03} + q_{30} + q_{31} + q_{32} + q_{33} + q_{60} + q_{61} + q_{62} + q_{63} \\
p_{0m.1} &= q_{10} + q_{11} + q_{14} + q_{15} + q_{40} + q_{41} + q_{44} + q_{45} + q_{70} + q_{71} + q_{74} + q_{75} \\
p_{01.1} &= q_{20} + q_{22} + q_{24} + q_{26} + q_{50} + q_{52} + q_{54} + q_{56} + q_{80} + q_{82} + q_{84} + q_{86}
\end{aligned}
$$

$$
\begin{aligned}
p_{10.1} &= q_{04} + q_{05} + q_{06} + q_{07} + q_{34} + q_{35} + q_{36} + q_{37} + q_{64} + q_{65} + q_{66} + q_{67} \\
p_{1m.1} &= q_{12} + q_{13} + q_{16} + q_{17} + q_{42} + q_{43} + q_{46} + q_{47} + q_{72} + q_{73} + q_{76} + q_{77} \\
p_{11.1} &= q_{21} + q_{23} + q_{25} + q_{27} + q_{51} + q_{53} + q_{55} + q_{57} + q_{81} + q_{83} + q_{85} + q_{87}
\end{aligned}
$$

which will be written in matrix form, $\vec{p} = \bar{P}\vec{q}$. This relationship between points in Q and P space imply the following constraints on points in P space:

$$
\begin{aligned}
p_{00.0} + p_{10.1} &\leq 1 \\
p_{0m.0} + p_{1m.1} &\leq 1 \\
p_{01.0} + p_{11.1} &\leq 1 \\
p_{10.0} + p_{00.1} &\leq 1 \\
p_{1m.0} + p_{0m.1} &\leq 1 \\
p_{11.0} + p_{01.1} &\leq 1
\end{aligned}
$$

Similar to the P space constraints in the binary case (Appendix A.1), we can prove that these constraints are necessary and sufficient for a point in P space to be modelled by some point in Q space.

ACE($D \rightarrow Y$) may be optimized given the constraints

$$
\begin{aligned}
\vec{p} &= \bar{P}\vec{q} \\
\sum_{j=0}^{8}\sum_{k=0}^{7} q_{jk} &= 1 \\
q_{jk} &\geq 0 \qquad j \in \{0,\ldots,8\} \text{ and } k \in \{0,\ldots,7\}
\end{aligned}
$$

using a program written for obtaining symbolic solutions to linear-programming problems. The following lower and upper bounds for ACE($D \rightarrow Y$) were obtained:

$$
L_{D\rightarrow Y}(\vec{p}) = \max \left\{ \begin{array}{c}
p_{00.0} + p_{11.1} - 1 \\
p_{00.1} + p_{11.1} - 1 \\
p_{11.0} + p_{00.1} - 1 \\
p_{00.0} + p_{11.0} - 1 \\
2p_{00.0} + p_{11.0} + p_{10.1} + p_{11.1} - 2 \\
p_{00.0} + 2p_{11.0} + p_{00.1} + p_{01.1} - 2 \\
p_{10.0} + p_{11.0} + 2p_{00.1} + p_{11.1} - 2 \\
p_{00.0} + p_{01.0} + p_{00.1} + 2p_{11.1} - 2 \\
-p_{0m.0} - p_{01.0} - p_{10.0} - p_{01.1} - p_{10.1} - p_{1m.1} \\
-p_{01.0} - p_{10.0} - p_{1m.0} - p_{0m.1} - p_{01.1} - p_{10.1}
\end{array} \right\}
$$

$$U_{D\to Y}(\vec{p}) = \min\left\{\begin{array}{c} 1-p_{10.0}-p_{01.1} \\ 1-p_{01.0}-p_{10.1} \\ 1-p_{01.0}-p_{10.0} \\ 1-p_{01.1}-p_{10.1} \\ 2-2p_{01.0}-p_{10.0}-p_{10.1}-p_{11.1} \\ 2-p_{01.0}-2p_{10.0}-p_{00.1}-p_{01.1} \\ 2-p_{10.0}-p_{11.0}-2p_{01.1}-p_{10.1} \\ 2-p_{00.0}-p_{01.0}-p_{01.1}-2p_{10.1} \\ p_{00.0}+p_{0m.0}+p_{11.0}+p_{00.1}+p_{1m.1}+p_{11.1} \\ p_{00.0}+p_{1m.0}+p_{11.0}+p_{00.1}+p_{0m.1}+p_{11.1} \end{array}\right\}$$

6.5 Conclusion

In this chapter, the strict bounds on the causal effect of treatment on observed response have been derived for models where received treatment take on non-binary values and treatment assignment and observed response are binary. These bounds can be used to derive useful bounds for treatment causal effects on quasi-experimental data containing continuous values of received treatment. By useful, we mean that a policy statement for treatment may be specified.

In future work, we might explore how exact knowledge of one point in the treatment-response curve can constrain the bounds of the rest of the treatment-response curve; for example, there might exist experimental data that has precisely determined the causal effect of a specific drug dosage on the probability of recovering from a particular ailment.

CHAPTER 7

Statistics in Law

7.1 Introduction

Over the last thirty years, there has been a steady increase in the number of court cases where statistical evidence has been introduced. These include cases involving product liability [BBB92], employment discrimination [MSZ84] [Fin80], price-fixing (anti-trust litigation) [DF85], genetic evidence, etc [Kay90]. Most of the statistical evidence provided is in the form of regression analysis [Fis80] [Fin80] [DF85], or by comparing the relative rates of some event across different populations, e.g., hiring rates among different races/sexes, or cases of illness among employees and non-employees. Most analysis of the data goes into showing the accuracy of the results given the sample size, but except for regression analysis, qualitative information is introduced after the statistical data has been presented.

The comparison of relative rates among different populations may produce serious errors in judgment, as will be shown by a hypothetical example in Section 7.2. These rates demonstrate dependence, but do not necessarily prove that those rates are a result of a defendants actions, because other unobserved factors may be responsible for the dependency. It turns out that the participants of the court case ask the right question (in terms of a counterfactual), e.g., "If the plaintiff were male, would she have been promoted a year earlier?"; however, although the right question is usually posed, the analysis usually fails to reflect the structure of the problem.

Ideally, a court would settle upon a qualitative model describing the causal relationships between variables in the system. Then the statistical data would be applied to parameterize the causal structure. Given this model of the system, the counterfactual conditional may then be evaluated and incorporated into the judgment.

When applying regression analysis, the variable claimed to be the basis of discrimination or unfair control is assumed to be exogenous to the system (i.e.,

a root node in the causal structure). Chapter 8 showed that not all variables that may be controlled satisfy this assumption, e.g., in a price-fixing case, the controlled variable is a product's price that is ordinarily influenced by other market factors.

It is very important that the model proposed by counsel is compatible with the statistical model provided as evidence. Such a case is discussed in [GKR94], where it was hypothesized that college attendance rates accounted for the difference in pass-rates between men and women on a test for promotion, yet the two distributions (one relating gender and college attendance, the other relating gender and rate of promotion) were inconsistent with a model where college attendance is supposed to explain away any gender bias. In a similar vein, Chapter 5 presented constraints (Eq. 5.13) on the observed distribution imposed by the assumed model of interaction in experimental studies with partial compliance. If a distribution fails these constraints, then the assumed model is improper for evaluating average treatment effects.

7.2 Hypothetical Product Safety Litigation

Evaluation of counterfactual probabilities could be enlightening in some legal cases in which a plaintiff claims that a defendant's actions were responsible for the plaintiff's misfortune. Improper rulings can easily be issued without an adequate treatment of counterfactuals. Consider the following hypothetical and fictitious case study, especially crafted to accentuate the disparity between different methods of analysis.

The marketer of PeptAid (antacid medication) randomly mailed out product samples to 10% of the households in the city of Stress, California. In a follow-up study, researchers determined for each individual whether they received the PeptAid sample, whether they consumed PeptAid, and whether they developed peptic ulcers in the following month.

The causal structure which describes the influences in this scenario is identical to the partial-compliance model given by Figure 5.1, where z_1 asserts that PeptAid was received from the marketer; d_1 asserts that PeptAid was consumed; and y_1 asserts that peptic ulceration occurred. The data showed the following distribution:

$$P(z_1) = 0.1$$

$$P(y_0, d_0|z_0) = 0.32 \qquad P(y_0, d_0|z_1) = 0.02$$
$$P(y_0, d_1|z_0) = 0.32 \qquad P(y_0, d_1|z_1) = 0.17$$
$$P(y_1, d_0|z_0) = 0.04 \qquad P(y_1, d_0|z_1) = 0.67$$
$$P(y_1, d_1|z_0) = 0.32 \qquad P(y_1, d_1|z_1) = 0.14$$

This data indicates a high-correlation between those individuals who consumed PeptAid and those who developed peptic ulcers in the following month

$$P(y_1|d_1) = 0.50 \qquad P(y_1|d_0) = 0.26$$

In addition, the intent-to-treat analysis showed that those individuals who received the PeptAid samples had a 45% greater chance of developing peptic ulcers

$$P(y_1|z_1) = 0.81 \qquad P(y_1|z_0) = 0.36$$

The plaintiff (Mr. Smith), having heard of the study, litigated against both the marketing firm and the PeptAid producer. The plaintiff's attorney argued against the producer, claiming that the consumption of PeptAid triggered his client's ulcer and resulting medical expenses. Likewise, the plaintiff's attorney argued against the marketer, claiming that his client would not have developed an ulcer, if the marketer had not distributed the product samples.

The defense attorney, representing both the manufacturer and marketer of PeptAid, though, rebutted this argument, stating that the high correlation between PeptAid consumption and ulcers was attributable to a common factor, namely, pre-ulcer discomfort. Individuals with gastrointestinal discomfort would be much more likely to both use PeptAid and develop stomach ulcers. To bolster his clients' claims, the defense attorney introduced expert analysis of the data showing that, on the average, consumption of PeptAid actually decreases an individual's chances of developing ulcers by at least 15%.

Indeed, the application of Eqs. 5.29 and 5.30 results in the following bounds on the average causal effect of PeptAid consumption on peptic ulceration

$$-0.23 \leq \text{ACE}(D \rightarrow Y) \leq -0.15$$

and proves that PeptAid is beneficial to the population as a whole.

The plaintiff's attorney, though, stressed the distinction between the average treatment effects for the entire population and the sub-population consisting of those individuals who, like his client, received the PeptAid sample, consumed it and then developed ulcers. Analysis of the population data indicated that had

PeptAid not been distributed, Mr. Smith would have had at most a 7% chance of developing ulcers regardless of any confounding factors such as pre-ulcer pain. Likewise, if Mr. Smith had not consumed PeptAid, he would have had at most a 7% chance of developing ulcers.

The damaging statistics against the marketer are obtained by evaluating the bounds on the probability that the plaintiff would have developed a peptic ulcer if he had not received the PeptAid sample, given that he in fact received the sample PeptAid, consumed the PeptAid, and developed peptic ulcers. This probability may be written in terms of the functional model parameters:

$$P(y_1^*|\hat{z}_0^*, y_1, d_1, z_1) = \frac{P(r_z{=}1)[q_{13} + q_{31} + q_{33}]}{P(y_1, d_1, z_1)}$$

But, since Z is a root node in the probabilistic specification, $P(r_z{=}1) = P(z_1)$; therefore,

$$\begin{aligned} P(y_1^*|\hat{z}_0^*, y_1, d_1, z_1) &= \frac{q_{13} + q_{31} + q_{33}}{P(y_1, d_1|z_1)} \\ &= \frac{q_{13} + q_{31} + q_{33}}{p_{11.1}}. \end{aligned}$$

This expression is linear with respect to the Q parameters; therefore, we may use linear optimization to derive symbolic bounds on the counterfactual probability with respect to the probabilistic specification $P(y, d|z)$:

$$\begin{aligned} &\frac{1}{p_{11.1}} \max \left\{ \begin{array}{c} 0 \\ p_{11.1} - p_{00.0} \\ p_{11.0} - p_{00.1} - p_{10.1} \\ p_{10.0} - p_{01.1} - p_{10.1} \end{array} \right\} \\ &\leq P(y_1^*|\hat{z}_0^*, z_1, d_1, y_1) \leq \\ &\frac{1}{p_{11.1}} \min \left\{ \begin{array}{c} p_{11.1} \\ p_{10.0} + p_{11.0} \\ 1 - p_{00.0} - p_{10.1} \end{array} \right\} \end{aligned}$$

Similarly, the damaging evidence against PeptAid's producer is obtained by evaluating the bounds on the counterfactual probability $P(y_1^*|\hat{d}_0^*, y_1, d_1, z_1)$. In terms of the Q parameters the counterfactual probability is written:

$$\begin{aligned} P(y_1^*|\hat{d}_0^*, y_1, d_1, z_1) &= \frac{q_{13} + q_{33}}{q_{11} + q_{13} + q_{31} + q_{33}} \\ &= \frac{q_{13} + q_{33}}{p_{11.1}}. \end{aligned}$$

If we minimize/maximize the numerator given the linear constraints, we arrive at the following bounds:

$$\frac{1}{p_{11.1}} \max \left\{ \begin{array}{c} 0 \\ p_{11.1} - p_{00.0} - p_{11.0} \\ p_{10.0} - p_{01.1} - p_{10.1} \end{array} \right\} \leq P(y_1^* | \hat{d}_0^*, z_1, d_1, y_1) \leq \frac{1}{p_{11.1}} \min \left\{ \begin{array}{c} p_{11.1} \\ p_{10.0} + p_{11.0} \\ 1 - p_{00.0} - p_{10.1} \end{array} \right\}$$

Substituting the observed distribution $P(y, d|z)$ into these formulas, the following bounds were obtained

$$0.00 \leq P(y_1^* | \hat{z}_0^*, z_1, d_1, y_1) \leq 0.07$$
$$0.00 \leq P(y_1^* | \hat{d}_0^*, z_1, d_1, y_1) \leq 0.07$$

We can write the average causal effects for the sub-population resembling the plaintiff by conditioning the counterfactual probabilities in Eqs. (5.16) and (5.17) on the features of the plaintiff.

$$\text{ACE}(D \rightarrow Y | z_1, d_1, y_1) = P(y_1^* | \hat{d}_1^*, z_1, d_1, y_1) - P(y_1^* | \hat{d}_0^*, z_1, d_1, y_1)$$

Counterfactual probabilities have the property that if the counterfactual antecedent is implied by the real-world observation, then the probability of the counterfactual consequent is the same as in the real-world given the observations:

$$P(c^* | \hat{a}^*, o) = P(c = c^* | o)$$

Therefore,

$$P(y_1^* | \hat{z}_1^*, z_1, d_1, y_1) = 1.00$$
$$P(y_1^* | \hat{d}_1^*, z_1, d_1, y_1) = 1.00$$

and

$$0.93 \leq \text{ACE}(D \rightarrow Y | z_1, d_1, y_1) \leq 1.00$$
$$0.93 \leq \text{ACE}(Z \rightarrow Y | z_1, d_1, y_1) \leq 1.00$$

At least 93% of the people in the plaintiff's subpopulation would not have developed ulcers had they not been encouraged to take PeptAid (z_0), or similarly, had they not taken PeptAid (d_0). This lends very strong support for the plaintiff's claim that he was adversely affected by the marketer and producer's actions and product.

The judge ruled in favor of the plaintiff. PeptAid withdrew the product from the market, and initiated a research effort to identify observable characteristics of those individuals who are adversely effected by PeptAid.

One might be curious about the distribution of consumption and response behaviors, $P(r_d, r_y)$, that would be responsible for such a peculiar story. Given the distribution over the observables $\{Z, D, Y\}$, we can evaluate bounds on each individual combination of consumption and response behaviors, leading to the identification of four common behaviors in the population:

$$0.16 \leq q_{10} \leq 0.17$$
$$0.13 \leq q_{11} \leq 0.14$$
$$0.31 \leq q_{22} \leq 0.32$$
$$0.31 \leq q_{23} \leq 0.32$$

Essentially, this tells us that about 1/3 of the population consists of individuals who would consume PeptAid (d_1) if and only if they received a sample in the mail (z_1). Of this sub-population (r_{d1}), about half of them would never develop ulcers (r_{y0}), while the other half would develop ulcers (y_1) if and only if they consumed PeptAid (r_{y1}).

The other 2/3 of the population consists of individuals who would consume PeptAid if and only if they do not receive a sample in the mail. Of this subpopulation (r_{d2}), about half of them develop ulcers if and only if they do not consume PeptAid (r_{y2}), while the other half would always develop ulcers (r_{y3}).

CHAPTER 8

Policy Analysis in Linear Models

8.1 Introduction

Counterfactual thinking dominates reasoning in political science and economics. We say, for example, "If Germany were not punished so severely at the end of World War I, Hitler would not have come to power," or "If Reagan did not lower taxes, our deficit would be lower today." Such thought experiments emphasize an understanding of generic laws in the domain and are aimed toward shaping future policy making, for example, "defeated countries should not be humiliated," or "lowering taxes (contrary to Reaganomics) tends to increase national debt."

Strangely, there is very little formal work on counterfactual reasoning or policy analysis in the behavioral science literature. An examination of a number of econometric journals and textbooks, for example, reveals an imbalance: while an enormous mathematical machinery is brought to bear on problems of estimation and prediction, policy analysis (which is the ultimate goal of economic theories) receives almost no formal treatment. Currently, the most popular methods driving economic policy making are based on so-called *reduced-form* analysis: to find the impact of a policy involving decision variables X on outcome variables Y, one examines past data and estimates the conditional expectation $E(Y|X=x)$, where x is the particular instantiation of X under the policy studied.

The assumption underlying this method is that the data were generated under circumstances in which the decision variables X act as exogenous variables, that is, variables whose values are determined outside the system under analysis. However, while new decisions should indeed be considered exogenous for the purpose of evaluation, past decisions are rarely enacted in an exogenous manner.[1]

[1]This distinction is often blurred in the literature. [DS93], for example, state: "A variable is considered exogenous to a system if its value is determined outside the system, either because we can control its value externally (e.g., the amount of taxes in a macro-economic model) or because we believe that this variable is controlled externally (like the weather in a system describing crop yields, market prices, etc.)" Still, our ability to externally control the value of a variable X does not render X exogenous for the purpose of legitimizing the reduced form

Almost every realistic policy (e.g., taxation) imposes control over some endogenous variables, that is, variables whose values are determined by other variables in the analysis. Let us take taxation policies as an example. Economic data are generated in a world in which the government is reacting to various indicators and various pressures; hence, taxation is endogenous in the data-analysis phase of the study. Taxation becomes exogenous when we wish to predict the impact of a specific decision to raise or lower taxes. The reduced-form method is valid only when past decisions are nonresponsive to other variables in the system, and this, unfortunately, eliminates most of the interesting control variables (e.g., tax rates, interest rates, quotas) from the analysis.[2]

This difficulty is not unique to economic or social policy making; it appears whenever one wishes to evaluate the merit of a plan on the basis of the past performance of other agents. Even when the signals triggering the past actions of those agents are known with certainty, a systematic method must be devised for selectively ignoring the influence of those signals from the evaluation process. In fact, the very essence of *evaluation* is having the freedom to imagine and compare trajectories in various counterfactual worlds, where each world or trajectory is created by a hypothetical implementation of a policy that is free of the very pressures that compelled the implementation of such policies in the past.

This chapter will present an example of counterfactual analysis in the area of econometrics, where apparently no adequate formalism for dealing with policy analysis has been proposed. In contrast to reduced-form analysis, our method allows evaluation of the consequences of intervening on economic attributes that are endogenous in normal operation only to become exogenous for the purpose

analysis: for $E[Y|X = x]$ to represent the impact of $X = x$ on Y, X must also be independent of all implicit factors (disturbance terms) affecting Y.

While every economist knows that this disturbance-independence is a necessary condition for consistent estimation of structural parameters, most economists assume that disturbance-independence is a guaranteed property of controllable policy variables. A popular textbook [Int78], for example, mentions these two properties as if they were synonymous: "The exogenous variables are variables the values for which are determined outside the model but which influence the model. From a formal standpoint the exogenous variables are assumed to be statistically independent of all stochastic disturbance terms of the model, while the endogenous variables are not statistically independent of those terms. ... In general the exogenous variables are either historically given, policy variables, or determined by some separate mechanism."

[2]This problem is unrelated to the celebrated Lucas's critique [Luc76] which concerns parameter changes due to economic agents becoming aware of interventions. The failure of reduced-form analysis extends to physical systems as well, where there are no rational agents to speak of, and where system parameters remain unaltered (except those under direct control).

of evaluation. The general techniques developed in Section 2.8.2 will be demonstrated in Section 8.2 by evaluating the effect on the demand for some commodity when a government imposes price controls on that commodity for the first time.

8.2 Example

Consider an econometric structural equation model described in [Gol92]

$$q = b_1 p + d_1 i + u_1 \tag{8.1}$$
$$p = b_2 q + d_2 w + u_2 \tag{8.2}$$

where, q = the quantity of household demand for product A, p = unit price of product A, i = household income, w = wage rate for producing product A, u_1 = demand shock, and u_2 = supply shock.

We extend this model by incorporating an additional variable r = household demand for some substitute product B, along with its structural equation

$$r = b_3 p + u_3$$

As an example, B could stand for tea and A for coffee.

Consider the following set of counterfactual queries:

1. What would be the expectation of demand for coffee (q) had we intervened to force coffee prices (p) to some predetermined value, say $p = 7$?

2. What would be the expectation of demand for coffee (q) had we intervened to force coffee prices (p) to some predetermined value, say $p = 7$, and then observed the demand for tea (r) to be $r = 4$?

3. Given that presently the demand for tea (r) is $r = 4$, what would be the expectation of demand for coffee (q) had we intervened to force coffee prices (p) to some predetermined value, say $p = 7$?

Note the difference between queries number 2 and 3. Number 2 states that the price intervention occurs prior to our observation of Product B's demand, while number 3 states that we first make an observation of Product B's demand and then intervene to force Product A's price.

The above counterfactual queries only involve the variables $\vec{X} = [P, Q, R]$; therefore, we may marginalize out all remaining variables in Eqs. (8.1) and (8.2),

only retaining the distributions on P, Q, and R's disturbance terms. Because I and W are exogenous (root) variables in the structural equations, we may combine I and U_1 into one disturbance variable ϵ_q. Likewise, W and U_2 may be combined into one disturbance variable ϵ_p. The structural equations for analyzing the above counterfactual queries may be reduced to

$$\begin{aligned} \vec{x} &= B\vec{x} + \vec{\epsilon} \\ \begin{bmatrix} p \\ q \\ r \end{bmatrix} &= \begin{bmatrix} 0 & b_2 & 0 \\ b_1 & 0 & 0 \\ b_3 & 0 & 0 \end{bmatrix} \begin{bmatrix} p \\ q \\ r \end{bmatrix} + \begin{bmatrix} \epsilon_p \\ \epsilon_q \\ \epsilon_r \end{bmatrix} \end{aligned} \tag{8.3}$$

The causal structure for this model is shown in Figure 8.1.

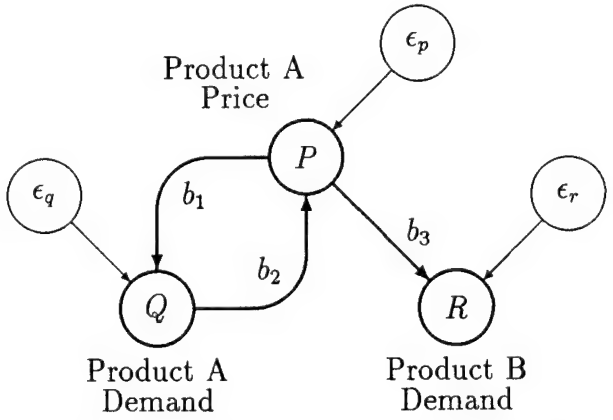

Figure 8.1: *Causal structure of an econometric model relating the demand for two products A and B and the price of product A. The variables are related according to the linear structural equations given in Eq. 8.3, where the disturbances, ϵ_p, ϵ_q, and ϵ_r are independent and normally distributed.*

Because R and Q are d-separated ([Pea88]) by P when the arrow $Q \longrightarrow P$ is removed, the observation of R after P's intervention has no impact on the evaluation of Q's distribution. Therefore, the counterfactual distribution of demand for coffee (Q) will be the same as evaluated from queries number 1 and 2.

Suppose that the parameters for this model are given by:

$$\begin{aligned} B &= \begin{bmatrix} 0 & 0.50 & 0 \\ -1.80 & 0 & 0 \\ 1.00 & 0 & 0 \end{bmatrix} \\ \vec{\mu}_\epsilon^t &= \begin{bmatrix} 0 & 19.00 & 3.00 \end{bmatrix} \end{aligned}$$

$$\Sigma_{\epsilon,\epsilon} = \begin{bmatrix} 1.00 & 0 & 0 \\ 0 & 3.00 & 0 \\ 0 & 0 & 2.00 \end{bmatrix}$$

which reflects the following prior distribution on $\vec{X} = [P, Q, R]$:

$$\vec{\mu}_x^t = \begin{bmatrix} 5.00 & 10.00 & 8.00 \end{bmatrix}$$
$$\Sigma_{x,x} = \begin{bmatrix} 0.48 & -0.08 & 0.48 \\ -0.08 & 1.73 & -0.08 \\ 0.48 & -0.08 & 2.48 \end{bmatrix}$$

The expected price of coffee is \$5.00, while the average demands for coffee and tea are 10 and 8 units, respectively.

The first query above is interested in determining the distribution of demand for coffee (Q) given that no observations have been made on the system, if we had intervened to force the price of coffee to \$7.00. Evaluating the expressions in Eqs. (2.14)–(2.15), we arrive at the following distribution:

$$\vec{\mu}_{x^*|\hat{p}=7}^t = \begin{bmatrix} 7.00 & 6.40 & 10.00 \end{bmatrix} \tag{8.4}$$
$$\sigma_{x^*|\hat{p}=7} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3.00 & 0 \\ 0 & 0 & 2.00 \end{bmatrix}$$

We conclude that the average household demand for coffee and tea would have been 6.4 and 10 units, respectively, if the price of coffee were \$7.00.

The third question asks what would have been the distribution of demand for coffee (Q), if the price of coffee were controlled to \$7.00, given that demand for tea is currently 4 units. Applying the expressions in Eqs. (2.14)–(2.15):

$$\vec{\mu}_{x^*|\hat{p}=7,r=4}^t = \begin{bmatrix} 7.00 & 5.13 & 6.78 \end{bmatrix} \tag{8.5}$$
$$\sigma_{x^*|\hat{p}=7,r=4} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 2.75 & -0.64 \\ 0 & -0.64 & 0.39 \end{bmatrix}$$

Note the importance of the observation of demand for tea (R). In the first query we found that forcing the price of coffee (P) to \$7.00 will reduce the expected demand for coffee (Q) from 10 units to 6.4 units. The observation of 4 unit demand for tea changes our belief in the expected value of the demand for coffee to

$\mu_{q|r=4} = 10.13$ units; if we intervene to force the price of coffee \$7.00, the expected demand for coffee (Q) will be reduced from 10.13 to 5.13 units. Therefore, we see that enforcing price control on coffee would have had a more adverse affect on the demand for coffee under the knowledge that the demand for tea was only 4 units. In addition, the expected household demand of tea would have been 6.78 units rather than the observed 4 units.

If we believe that the disturbance on the demand for coffee (ϵ_q) slowly change, or at least change infrequently, then we can use the results of this counterfactual distribution to determine whether price controls should now be imposed to meet our needs. In other words, the counterfactual distribution will tell us how we expect variables' distributions to change as a result of an external intervention applied in the present.

It is important to note the difference between counterfactual distributions (conditioned on observations and external intervention) and distributions simply conditioned on observations. Consider the distribution that would be computed from observing the price of coffee at \$7.00 ($p = 7$), or from observing the demand for tea at 4 units and the coffee price at \$7.00 ($r = 4$ and $p = 7$):

$$\vec{\mu}^t_{x|p=7} = \begin{bmatrix} 7.00 & 9.66 & 10.00 \end{bmatrix} \tag{8.6}$$

$$\sigma_{x,x|p=7} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1.71 & 0 \\ 0 & 0 & 2.00 \end{bmatrix} \tag{8.7}$$

$$\vec{\mu}^t_{x|r=4,p=7} = \begin{bmatrix} 7.00 & 9.66 & 4.00 \end{bmatrix} \tag{8.8}$$

$$\sigma_{x,x|r=4,p=7} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1.71 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{8.9}$$

Contrast the expected household demand for coffee evaluated from these conditional distributions to those counterfactual distributions where the price of coffee (P) has been forced by external intervention. In particular, compare Eq. (8.6) to Eq. (8.4) and Eq. (8.8) to Eq. (8.5). This should convince the reader that it is incorrect to use distributions conditioned on observations for evaluating (economic) policies, because they fail to capture the change in value of the variable that will undergo external intervention. The expected value of that variable prior to intervention is important for properly evaluating the effect of that intervention.

8.3 Conclusion

This chapter has addressed the inadequacy of current techniques in econometrics and the social sciences for evaluating the potential effects of economic and social policies. Current techniques fail to correctly evaluate policies that control endogenous variables, that is, variables that are influenced by other variables in the system prior to enacting the policy.

This deficiency has been addressed by applying the formalism for evaluating counterfactual conditionals in linear structural equation models described in Section 2.8.2. This method is applicable to the analysis of policies, even when the policy dictates intervention on an endogenous variable. An example was presented that demonstrates the disparity between analyses based on counterfactuals and reduced-form analysis which treats intervention as an observation on controlled variables.

CHAPTER 9

Conclusion

Counterfactual reasoning is common to everyday discourse, and is important to a broad range of applications including liability litigation and policy analysis. Although closest-world semantics and imaging provide a solid foundation for analyzing counterfactual conditionals, a formalization of closeness of worlds that intuitively reflects our understanding of the mechanisms that drive the world has not previously been provided. This dissertation has addressed this short-coming by representing generic knowledge of the world by causal relationships and by interpreting a counterfactual antecedent as an external intervention that forces the antecedent to be true despite all known influences that normally impinge on the antecedent variable.

Under this formulation, the ability to precisely evaluate a counterfactual probability (i.e., the probability that the consequent would have been true, if the antecedent were true) is dependent on the detail of causal knowledge available. While a counterfactual probability may be uniquely computed given a functional model of a system, only bounds on the counterfactual probability may be computed if the causal relationships are parameterized by a probabilistic specification (i.e., a conditional probability distribution for each variable given an instantiation of its causal influences). Depending on the form of a counterfactual query and the causal structure of the system, it is not always possible to guarantee the evaluation of bounds on a counterfactual probability. However, it has been shown in this dissertation that the evaluation of bounds is guaranteed for counterfactual beliefs when the causal model is parameterized by order-of-magnitude probabilities.

Our formulation for interpreting and evaluating counterfactual probabilities has been applied to the determination of bounds on treatment effects from studies in which subject compliance is imperfect, resulting in tighter bounds than previously discovered. These results are based on a large-sample approximation; in the future, we would like to explore the small-sample analysis through hypothesis testing and computing the distribution of bounds.

We have also demonstrated the potential power of counterfactual probabilities for determining liability in legal cases when a causal formulation may be brought to bear in the case. Finally, economic and social policy analysis can also benefit from evaluating counterfactuals in structural equation models, which allows analysts to determine the effect of controlling endogenous variables in a system.

APPENDIX A

Proofs

A.1 Sufficiency of P space constraints

In this section, we will prove that any distribution in observation space, P, which satisfies the constraints given in equation 5.13 can be modeled by the latent structure given in figure 5.1.

The key to the proof is to show that there is a one-to-one mapping between the extreme points in the observation space constrained by equation 5.13 and the extreme points in a transformed parameter space of the counterfactual model. This can easily be accomplished by using an algorithm for enumerating all vertices in a convex polytope.

Theorem A.1.1 *[Sufficiency of P space constraints] Satisfaction of the constraints:*

$$\begin{aligned} p_{11.1} + p_{01.0} &\leq 1 \\ p_{01.1} + p_{11.0} &\leq 1 \\ p_{10.1} + p_{00.0} &\leq 1 \\ p_{00.1} + p_{10.0} &\leq 1 \end{aligned}$$

is sufficient to guarantee that the latent structure in figure 5.1 can model a point in the probabilistic observation space $\vec{p}$.

Proof:

The full set of linear constraints (including those imposed by probability theory) which define the above P space is given by

$$\begin{aligned} p_{11.1} + p_{01.0} &\leq 1 \\ p_{01.1} + p_{11.0} &\leq 1 \\ p_{10.1} + p_{00.0} &\leq 1 \end{aligned}$$

$$
\begin{aligned}
p_{00.1} + p_{10.0} &\leq 1 \\
p_{00.1} + p_{01.1} + p_{10.1} + p_{11.1} &= 1 \\
p_{00.0} + p_{01.0} + p_{10.0} + p_{11.0} &= 1 \\
p_{00.0}, p_{01.0}, p_{10.0}, p_{11.0}, p_{00.1}, p_{01.1}, p_{10.1}, p_{11.1} &\geq 0
\end{aligned}
$$

The extreme vertices within this closed polytope may be enumerated by one of many vertex enumeration algorithms (for example, [Mat73])

$$
\begin{aligned}
\vec{p}_1 &= (1,0,0,0,1,0,0,0) \\
\vec{p}_2 &= (1,0,0,0,0,1,0,0) \\
\vec{p}_3 &= (1,0,0,0,0,0,0,1) \\
\vec{p}_4 &= (0,1,0,0,1,0,0,0) \\
\vec{p}_5 &= (0,1,0,0,0,1,0,0) \\
\vec{p}_6 &= (0,1,0,0,0,0,1,0) \\
\vec{p}_7 &= (0,0,1,0,0,1,0,0) \\
\vec{p}_8 &= (0,0,1,0,0,0,1,0) \\
\vec{p}_9 &= (0,0,1,0,0,0,0,1) \\
\vec{p}_{10} &= (0,0,0,1,1,0,0,0) \\
\vec{p}_{11} &= (0,0,0,1,0,0,1,0) \\
\vec{p}_{12} &= (0,0,0,1,0,0,0,1)
\end{aligned}
$$

where

$$
\vec{p} = (p_{00.0}, p_{01.0}, p_{10.0}, p_{11.0}, p_{00.1}, p_{01.1}, p_{10.1}, p_{11.1})
$$

The transformation from Q space to P space (Eq. 5.25) was explicated in Section 5.2.2. There are four pairs of Q space parameters each of which always occur in combination within this transformation; therefore we can reduce the Q space to a 12 dimensional space, V, where V and Q are related as follows:

$$
\begin{aligned}
v_1 &= q_{00} + q_{01} \\
v_2 &= q_{02} + q_{03} \\
v_3 &= q_{10} \\
v_4 &= q_{11} \\
v_5 &= q_{12}
\end{aligned}
$$

$$
\begin{aligned}
v_6 &= q_{13} \\
v_7 &= q_{20} \\
v_8 &= q_{21} \\
v_9 &= q_{22} \\
v_{10} &= q_{23} \\
v_{11} &= q_{30} + q_{32} \\
v_{12} &= q_{31} + q_{33}.
\end{aligned}
$$

The V and P spaces are then related by the following equations:

$$
\begin{aligned}
p_{00.0} &= v_1 + v_3 + v_4 \\
p_{01.0} &= v_7 + v_9 + v_{11} \\
p_{10.0} &= v_2 + v_5 + v_6 \\
p_{11.0} &= v_8 + v_{10} + v_{12}
\end{aligned}
\tag{A.1}
$$

$$
\begin{aligned}
p_{00.1} &= v_1 + v_7 + v_8 \\
p_{01.1} &= v_3 + v_5 + v_{11} \\
p_{10.1} &= v_2 + v_9 + v_{10} \\
p_{11.1} &= v_4 + v_6 + v_{12}
\end{aligned}
\tag{A.2}
$$

If we constrain V by probability theory ($\sum_{i=1}^{12} v_i = 1$ and $v_i \geq 0, i = 1 \ldots 12$) we obtain twelve extreme vertices corresponding to the points where exactly one of the v_i is equal to 1.0 and all others are zero.

Eq. A.2 provides a one-to-one mapping between these twelve V space vertices and the twelve vertices in the constrained P space. Because the linear transformation maps the extreme vertices of the V space to the extreme vertices of the P space, then for every point $\vec{p}$ in the constrained P space, there exists a point in V (and hence Q) space that models $\vec{p}$. Since the latent structure model (Figure 5.1) subsumes the response-function model, any point in P which satisfies the constraints given by equation 5.13 can be modeled using the latent structure.

□

APPENDIX B

Closed-form solutions to linear optimization

In general, a linear optimization problem may be specified by an objective function to minimize

$$\min c^t x \tag{B.1}$$

along with a set of linear constraints that must be satisfied:

$$Ax \geq b \tag{B.2}$$

$$x \geq 0 \tag{B.3}$$

where A is a matrix of coefficients, c is a vector of coefficients acting on the variable vector x, and b is a vector of constants. Given A, c, and b, there are many algorithms that will return a value for the vector x that globally minimizes $c^t x$ subject to the specified linear constraints [Had62] [Dan63].

Sometimes, though, it is desirable to derive a closed-form expression for $\min c^t x$ for all possible values of the constraint vector b. The procedure for deriving this closed-form solution is tied to the enumeration of all extreme vertices in the dual linear-programming problem.

The dual of the above minimization problem is given by the objective function:

$$\max y^t b$$

subject to the constraints

$$y^t A \leq c^t \tag{B.4}$$

$$y \geq 0 \tag{B.5}$$

It is known [Jac93] that the expression for the $\min c^t x$ in terms of b is given by

$$\min c^t x = \max_{i=1,\ldots,K} \bar{y}_i^t b \tag{B.6}$$

where $\{\bar{y}_i | i = 0, \ldots, K\}$ is the set of y such that each $\bar{y}_i$ maximizes $y^t b$ for some value of b. This set of y is exactly the set of extreme vertices in the constraint space given by Eqs. (B.4) and (B.5)

$$\begin{aligned} y^t A &\leq c^t \\ y &\geq 0 \end{aligned}$$

Therefore, to generate the general solution to the minimization problem given by Eqs. (B.1)-(B.3) we merely enumerate all vertices in the constraint space of the dual linear-programming problem (Eqs. (B.4) and (B.5)) and substitute into Eq. (B.6).

A review of some vertex-enumeration algorithms may be found in [MR80], of which the algorithm by Mattheiss [Mat73] was implemented to derive the solutions presented in this dissertation.

In order to apply this procedure for deriving a closed-form solution to a linear-optimization problem, the constraints must be transformed to $\geq$ relations. Many of the constraints imposed for deriving bounds on counterfactual probabilities are in the form of equalities. The presence of equality constraints indicates that there are fewer degrees of freedom in the problem space than the number of variables suggests; these equalities will be used to eliminate variables from the linear-optimization problem. For example, suppose that the following constraints exist for the two variables a and b:

$$\begin{aligned} a + b &= 1 \\ a &\geq 0 \\ b &\geq 0 \end{aligned}$$

and the expression to be optimized is

$$2a - b$$

The equality relation allows us to write a in terms of the remaining variable

$$a = 1 - b$$

which may then be used to eliminate a from the linear programming problem. The constraints then become

$$\begin{aligned} 1 - b &\geq 0 \\ b &\geq 0 \end{aligned}$$

while the objective function becomes

$$2 - 3b$$

B.1 Example

Consider the derivation of bounds for average treatment effect in Section 5.2.2; the objective function to optimize was given by Eq. (5.23) and the linear constraints were given by Eq. (5.26).

The equality constraints in this specification allow us to eliminate seven of the variables $q_{00}, q_{10}, q_{20}, q_{11}, q_{21}, q_{02}, q_{12}$ resulting in the following seven non-trivial inequality constraints:

$$\begin{aligned}
q_{30} - q_{01} + q_{31} + q_{22} + q_{32} + q_{23} + q_{33} &\geq p_{01.0} + p_{11.0} - p_{00.1} \\
-q_{30} - q_{22} - q_{32} + q_{13} - q_{23} &\geq p_{10.0} - p_{01.1} - p_{10.1} \\
-q_{30} - q_{22} - q_{32} &\geq -p_{01.0} \\
-q_{31} - q_{13} - q_{33} &\geq -p_{11.1} \\
-q_{31} - q_{23} - q_{33} &\geq -p_{11.0} \\
-q_{22} - q_{03} - q_{23} &\geq -p_{10.1} \\
q_{22} - q_{13} + q_{23} &\geq p_{10.1} - p_{10.0}
\end{aligned}$$

The objective function to be minimized $\text{ACE}(D \to Y)$ may also be rewritten in terms of the remaining variables:

$$p_{11.0} + p_{11.1} - p_{10.0} + q_{01} - q_{31} - q_{22} - q_{32} + q_{03} - q_{23} - 2q_{33}$$

Before this expression is minimized the constant terms $(p_{11.0} + p_{11.1} - p_{10.0})$ will be dropped. After the minimization is complete these terms will be reattached. Therefore, the expression to be optimized by linear programming is given by:

$$q_{01} - q_{31} - q_{22} - q_{32} + q_{03} - q_{23} - 2q_{33}$$

In terms of Eqs: (B.1)–(B.5), this task may be specified by the following matrices.

$$A = \begin{bmatrix}
1 & -1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\
-1 & 0 & 0 & -1 & -1 & 0 & 1 & -1 & 0 \\
-1 & 0 & 0 & -1 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & -1 \\
0 & 0 & -1 & 0 & 0 & 0 & 0 & -1 & -1 \\
0 & 0 & 0 & -1 & 0 & -1 & 0 & -1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & -1 & 1 & 0
\end{bmatrix}$$

$$
\begin{aligned}
b &= \begin{bmatrix} p_{01.0} + p_{11.0} - p_{00.1} \\ p_{10.0} - p_{01.1} - p_{10.1} \\ -p_{01.0} \\ -p_{11.1} \\ -p_{11.0} \\ -p_{10.1} \\ p_{10.1} - p_{10.0} \end{bmatrix} \\
x^t &= \begin{bmatrix} q_{30} & q_{01} & q_{31} & q_{22} & q_{32} & q_{03} & q_{13} & q_{23} & q_{33} \end{bmatrix} \\
c^t &= \begin{bmatrix} 0 & 1 & -1 & -1 & -1 & 1 & 0 & -1 & -2 \end{bmatrix} \\
y^t &= \begin{bmatrix} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 \end{bmatrix}
\end{aligned}
$$

Applying a vertex enumeration algorithm to the dual linear-programming problem's constraint space leads to the following list of extreme vertices:

$$
\begin{aligned}
\bar{y}_1^t &= [\ 0\ 0\ 1\ 2\ 0\ 1\ 0\ ] \\
\bar{y}_2^t &= [\ 0\ 0\ 1\ 0\ 2\ 0\ 0\ ] \\
\bar{y}_3^t &= [\ 0\ 1\ 0\ 0\ 2\ 1\ 1\ ] \\
\bar{y}_4^t &= [\ 0\ 1\ 0\ 2\ 0\ 0\ 0\ ] \\
\bar{y}_5^t &= [\ 0\ 1\ 0\ 1\ 1\ 0\ 0\ ] \\
\bar{y}_6^t &= [\ 0\ 0\ 1\ 1\ 1\ 0\ 0\ ] \\
\bar{y}_7^t &= [\ 0\ 2\ 0\ 2\ 0\ 0\ 0\ ] \\
\bar{y}_8^t &= [\ 0\ 0\ 2\ 0\ 2\ 0\ 1\ ]
\end{aligned}
$$

Substituting these vertices into Eq. (B.6) and then adding back the previously dropped terms ($p_{11.0} + p_{11.1} - p_{10.0}$) will produce the same results given by Eq. (5.29).

B.2 Program Implementation

At the UCLA Cognitive Systems Laboratory, we have implemented a program for deriving closed-form solutions to these linear optimization problems. This program accepts a text file specifying the optimization problem to solve, and enumerates all terms in the minimization (maximization) set that make up the close-form solution for the maximum (minimum) of the specified objective function.

B.2.1 Input Text File

The input text file is composed of several sections (the order being irrelevant), each delimited by a header. Blank lines may be used to separate lines in the text file. In addition, comments may be entered on lines by placing a single '%' before free-form text. If an equation or expression is too long to fit on a single line, you may break the line by placing a backslash ('\') at the end of the line, and continuing the expression on the following line. This may be repeated to extend an expression over several lines. Symbol (variable or parameter) names must begin with an alphabetic character followed by alphabetic, numeric, or underscore characters (e.g., *a_01*, *Long_Name_32*) and must not exceed 20 characters in length. The following paragraphs explain the format of each section in the text file.

VARIABLES This section lists the variables that correspond to the vector x in Eq. (B.1). When bounding counterfactual probabilities, the values of these variables specify the distribution of response-functions, e.g., the Q space parameters in Chapter 5. Only one variable may be listed per line, and the *VARIABLES* keyword must appear on a line by itself.

PARAMETERS This section lists the parameters that correspond to the vector b in Eq. (B.2). When bounding counterfactual probabilities, these correspond to the conditional probability distributions over the observable variables, e.g., the P space parameters in Chapter 5. Only one parameter may be listed per line, and the *PARAMETERS* keyword must appear on a line by itself.

CONSTRAINTS This section lists the constraints imposed on the variables by the parameters. These constraints may be written as $=$, $\geq$, or $\leq$ relations. The constraints must be linear, i.e., plus, minus, and real coefficients. There is no requirement as to the placement of variables versus parameters or additive constants. Suppose that the set of variables is given by $\{X, Y, Z\}$, and the set of parameters is given by $\{A, B, C\}$, then the following constraints satisfy the required format:

$$\begin{aligned} -A + 8.9X - 4.3B &\leq C + 4.678 \\ A + B &= 0.5X \\ X + Y &\geq 1 + Z \end{aligned}$$

Nonnegativity constraints are assumed for all variables, e.g.,

$$\begin{aligned} X &\geq 0 \\ Y &\geq 0 \end{aligned}$$

Only one constraint may be listed per line (although it may extend over several lines using a backslash at the end of each incomplete line), and the *CONSTRAINTS* keyword must appear on a line by itself.

MINIMIZE/MAXIMIZE This keyword indicates whether the objective function is to be minimized or maximized, respectively. This keyword must appear on a line by itself.

OBJECTIVE The objective to be optimized must be an expression that is a linear function of the parameters, variables, and real constants. For example,

$$2A + X + 0.5Y - 6.7$$

END The specification of the optimization problem must be terminated by the *END* keyword alone on a separate line.

An example of a complete input file used for obtaining the results in Chapter 5 follows:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% This linear optimization specification will be used
% to generate the lower bounds on the average treatment
% effect for experimental studies where subject compliance
% is not perfect.
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
VARIABLES
    q00
    q01
    q02
    q03

    q10
    q11
    q12
```

```
    q13

    q20
    q21
    q22
    q23

    q30
    q31
    q32
    q33

PARAMETERS
    p00_0
    p01_0
    p10_0
    p11_0

    p00_1
    p01_1
    p10_1
    p11_1

CONSTRAINTS
    1 = q00 + q_01 + q_02 + q_03 + \
        q10 + q_11 + q_12 + q_13 + \
        q20 + q_21 + q_22 + q_23 + \
        q30 + q_31 + q_32 + q_33

    p00_0 = q00 + q01 + q10 + q11
    p01_0 = q20 + q22 + q30 + q32
    p10_0 = q02 + q03 + q12 + q13
    p11_0 = q21 + q23 + q31 + q33

    p00_1 = q00 + q01 + q20 + q21
    p01_1 = q10 + q12 + q30 + q32
    p10_1 = q02 + q03 + q22 + q23
    p11_1 = q11 + q13 + q31 + q33
```

```
MINIMIZE

OBJECTIVE
    q01 + q11 + q21 + q31 - q02 - q12 - q22 - q32

END
```

B.2.2 Program Output

The output of the program will first redisplay the problem specification in a canonical form. It will then display a set of expressions that are to be minimized or maximized depending on whether the objective function was to be maximized or minimized, respectively. These expressions will be linear functions of the specification parameters (not the variables) and a real constant.

For example, the output to the specification file shown above will be:

```
Constraints:
    q00 + q_01 + q_02 + q_03 + q10 + q_11 + q_12 + q_13 + q20 + \
q_21 + q_22 + q_23 + q30 + q_31 + q_32 + q_33 - 1 = 0
    p00_0 - q00 - q01 - q10 - q11 = 0
    p01_0 - q20 - q22 - q30 - q32 = 0
    p10_0 - q02 - q03 - q12 - q13 = 0
    p11_0 - q21 - q23 - q31 - q33 = 0
    p00_1 - q00 - q01 - q20 - q21 = 0
    p01_1 - q10 - q12 - q30 - q32 = 0
    p10_1 - q02 - q03 - q22 - q23 = 0
    p11_1 - q11 - q13 - q31 - q33 = 0

Minimize Objective:
    q01 + q11 + q21 + q31 - q02 - q12 - q22 - q32

Solution:
    MAX
    {
        p111 + p000 - 1,
        p110 + p001 - 1,
        - p011 - p101,
```

```
    - p010 - p100,
    p110 - p111 - p101 - p010 - p100,
    p111 - p110 - p100 - p011 - p101,
    p001 - p011 - p101 - p010 - p000,
    p000 - p010 - p100 - p011 - p001
}
```

References

[AI92] Joshua D. Angrist and Guido W. Imbens. "Average Causal Response with Variable Treatment Intensity." Discussion Paper 9234, Center for Economic Research, Tilburg University, October 1992.

[AIR93] J.D. Angrist, G.W. Imbens, and D.B. Rubin. "Identification of Causal Effects Using Instrumental Variables." Technical Report No. 136, Department of Economics, Harvard University, Cambridge, MA, June 1993.

[BBB92] Vincent M. Brannigan, Vicki M. Bier, and Christine Berg. "Risk, statistical inference, and the law of evidence: The use of epidemiological data in toxic tort cases." *Risk Analysis*, **12**(3):343–351, 1992.

[Bou92] Craig Boutilier. "A Logic for Revision and Subjunctive Queries." In *Proceedings Tenth National Conference on Artificial Intelligence*, pp. 609–15, Menlo Park, CA, 1992. AAAI Press.

[BP93] Alexander Balke and Judea Pearl. "Nonparametric bounds on causal effects from partial compliance data." Technical Report R-199, Cognitive Systems Laboratory, Computer Science Department, UCLA, September 1993. Submitted.

[BT84] Roger J. Bowden and Darrell A. Turkington. *Instrumental Variables.* Cambridge University Press, Cambridge, MA, 1984.

[cou93] J.-J.Ch. Meyer and W. van der Hoek. "Counterfactual reasoning by (means of) defaults." *Annals of Mathematics and Artificial Intelligence*, **9**:345–360, 1993.

[Dal88] M. Dalal. "Investigations into a theory of knowledge base revision: Prel iminary report." In *Proceedings of the Seventh National Conference on Artificial Intelligence*, pp. 475–479, 1988.

[Dan63] George B. Dantzig. *Linear Programming and Extensions.* Princeton University Press, Princeton, NJ, 1963.

[Dem69] Arthur P. Dempster. *Elements of Continuous Multivariate Analysis.* Addison-Wesley Publishing Company, 1969.

[DF85] R.S. Daggett and D.A. Freedman. "Econometrics and the Law: A Case Study in the Proof of Antitrust Damages." In L.M. Le Cam and R.A. Olshen, editors, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume I. Wadsworth, Inc., 1985.

[DM81] K. Roscoe Davis and Patrick G. McKeown. *Quantitative Models for Management.* Kent Publishing Company, Boston, MA, 1981.

[DS93] Marek J. Druzdzel and Herbert A. Simon. "Causality in Bayesian Belief Networks." In *Proceedings of the 9th Annual Conference on Uncertainty in Artificial Intelligence (UAI-93)*, pp. 3–11, 1993.

[EF91] B. Efron and D. Feldman. "Compliance as an Explanatory Variable in Clinical Trials." *Journal of the American Statistical Association*, **86**(413):9–26, March 1991.

[Fin80] Michael O. Finkelstein. "The judicial reception of multiple regression studies in race and sex descrimination cases." *Columbia Law Review*, **80**:737–754, 1980.

[Fis80] Franklin M. Fisher. "Multiple regression in legal proceedings." *Columbia Law Review*, **80**:702–736, 1980.

[Gin86] Matthew L. Ginsberg. "Counterfactuals." *Artificial Intelligence*, **30**:35–79, 1986.

[GKR94] Joseph Gastwirth, Abba Krieger, and Paul Rosenbaum. "How a Court Accepted an Impossible Explanation." *The American Statistician*, **48**(4):313–315, 1994.

[GM94] Gösta Grahne and Alberto O. Mendelzon. "Updates and subjunctive queries." In R. Demolombe and T. Imielinski, editors, *Nonstandard Queries and Nonstandard Answers*, chapter 10, pp. 255–279. Clarendon Press, Oxford, 1994.

[Gol92] Arthur S. Goldberger. "Models of substance; comment on N. Wermuth, 'On block-recursive linear regression equations'." *Brazilian Journal of Probability and Statistics*, **6**:1–56, 1992.

[Goo83] Nelson Goodman. *Fact, Fiction, and Forecast.* Harvard University Press, Cambridge, MA, 4 edition, 1983.

[Gra91] Gösta Grahne. "Updates and Counterfactuals." In J. Allen, R. Fikes, and E. Sandewall, editors, *Principles of Knowledge Representation and Reasoning*, pp. 269–276, San Mateo, CA, 1991. Morgan Kaufmann.

[Had62] G. Hadley. *Linear Programing.* Addison-Wesley, Reading, MA, 1962.

[Hec93] David Heckerman. "Causal Independence for Knowledge Acquisition and Inference." In Heckerman and Mamdani, editors, *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 122–127, San Mateo, CA, 1993. Morgan Kaufmann.

[Hol88] Paul W. Holland. "Causal Inference, Path Analysis, and Recursive Structural Equations Models." In C. Clogg, editor, *Sociological Methodology*, pp. 449–484. American Sociological Association, Washington, DC, 1988.

[HSP81] William L. Harper, Robert Stalnaker, and Glenn Pearce, editors. *Ifs: Conditionals, Belief, Decision, Chance, and Time.* D.Reidel Publishing Co., Boston, MA, 1981.

[Int78] Michael D. Intriligator. *Econometric models, techniques, and applications.* Prentice-Hall, 1978.

[Jac89] Peter Jackson. "On the Semantics of Counterfactuals." In N.S. Sridharan, editor, *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, pp. 1382–7 vol. 2, Palo Alto, CA, 1989. Morgan Kaurmann.

[Jac93] S.E. Jacobsen. "Class Notes on Linear Optimization (EE236A)." Electrical Engineering Deparment, UCLA, Fall 1993.

[Kay90] David H. Kaye. "Improving legal statistics." *Law and Society Review*, **24**(5):1255–1275, 1990.

[Lew76] David Lewis. "Probability of Conditionals and Conditional Probabilities." *The Philosophical Review*, **85**:297–315, 1976.

[Lew79] David Lewis. "Counterfactual Dependence and Time's Arrow." *Noûs*, pp. 455–476, 1979.

[Luc76] R.E. Lucas. "Economic policy evaluation: A critique." In K. Brunner and A.H. Meltzer, editors, *The Phillips Curve and Labor Markets*, pp. 19–46. North-Holland, 1976.

[Man90] Charles F. Manski. "Nonparametric Bounds on Treatment Effects." *American Economic Review, Papers and Proceedings*, **80**:319–323, May 1990.

[Mat73] T.H. Mattheiss. "An algorithm for determining irrelevant constraints and all vertices in systems of linear inequalities." *Operations Research*, **21**:247–260, 1973.

[MR80] T.H. Mattheiss and David S. Rubin. "A survey and comparison of methods for finding all vertices of convex polyhedral sets." *Mathematics of Operations Research*, **5**(2):167–185, May 1980.

[MSZ84] Paul Meier, Jerome Sacks, and Sandy L. Zabell. "What happened in Hazelwood: Statistics, employment descrimination, and the 80% rule." *American Bar Foundation Research Journal*, pp. 139–186, 1984.

[Nut80] Donald Nute. *Topics in Conditional Logic.* D. Reidel Publishing Company, Boston, 1980.

[PAA91] Luis M. Pereira, Joaquim N. Aparicio, and Jose J. Alferes. "Counterfactual Reasoning Based on Revising Assumptions." In V. Saraswat and K. Ueda, editors, *Logic Programming. Proceedings of the 1991 International Symposium*, pp. 566–577, Cambridge, MA, 1991. MIT Press.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufman Publishers, San Mateo, CA, 1988.

[Pea93a] Judea Pearl. "Aspects of Graphical Models Connected with Causality." In *Proceedings ofthe 49th Session of the International Statistical Institute*, pp. 391–401, Florence, Italy, August 1993. Short version in Statistical Science.

[Pea93b] Judea Pearl. "Aspects of Graphical Models Connected with Causality." Technical Report R-195-LL Revision II, UCLA Cognitive Systems Laboratory, June 1993. Presented at the *49th Session of the International Statistical Institute*, Florence, Italy, August 25 - September 3, 1993.

[Pea93c] Judea Pearl. "From Adams' Conditionals to Default Expressions, Causal Conditionals, and Counterfactuals." Technical Report R-193,

UCLA Cognitive Systems Laboratory, February 1993. To appear in *Festschrift for Ernest Adams*, Cambridge University Press, 1994.

[Pea93d] Judea Pearl. "From Conditional Oughts to Qualitative Decision Theory." In David Heckerman and Abe Mamdani, editors, *Uncertainty in Artificial Intelligence, Proceedings of the Ninth Conference*, pp. 12–20. Morgan Kaufmann, 1993.

[Pea94a] Judea Pearl. "A Probabilistic Calculus of Actions." In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, WA*, pp. 454–462, San Fransisco, CA, 1994. Morgan Kauffman.

[Pea94b] Judea Pearl. "A Probabilistic Calculus of Actions." Technical Report R-212, UCLA Cognitive Systems Laboratory, 1994. This volume (UAI-94).

[Poo93] David Poole. "Probabilistic Horn abduction and Bayesian networks." *Artificial Intelligence*, **64**(1):81–130, 1993.

[Pro84] Lipid Research Clinic Program. "The Lipid Research Clinics Coronary Primary Prevention Trial Results, Parts I and II." *Journal of the American Medical Association*, **251**(3):351–374, January 1984.

[PV91] Judea Pearl and Thomas Verma. "A Theory of Inferred Causation." In James Allen, Richard Fikes, and Erik Sandewall, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, pp. 441–452. Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[Rob89] J.M. Robins. "The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies." In L. Sechrest, H. Freeman, and A. Mulley, editors, *Health Service Research Methodology: A Focus on AIDS*, pp. 113–159. NCHSR, U.S. Public Health Service, 1989.

[RR83] P. Rosenbaum and D. Rubin. "The central role of propensity score in observational studies for causal effects." *Biometrika*, **70**:41–55, 1983.

[Rub74] Donald B. Rubin. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology*, **66**(5):688–701, 1974.

[Sca85] L.E. Scales. *Introduction to Non-Linear Optimization.* Springer-Verlag, New York, 1985.

[SGS91] Peter Spirtes, Clark Glymour, and Richard Scheines. "From Probability to Causality." *Philosophical Studies*, **64**(1):1–36, October 1991.

[SGS93] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* Springer, New York, 1993.

[Sky80] Brian Skyrms. "The Prior Propensity Account of Subjunctive Conditionals." In W.L. Harper, R. Stalnaker, and G. Pearce, editors, *Ifs*, pp. 259–265. D. Reidel Publishing Company, 1980.

[Spo88] Wolfgang Spohn. "Ordinal Conditional Functions: A Dynamic Theory of Epistemic States." In William L. Harper and Brian Skyrms, editors, *Causation in Decision, Belief Change, and Statistics, II*, pp. 105–134. Kluwer Academic Publishers, Boston, 1988.

[SR66] H.A. Simon and N. Rescher. "Cause and Counterfactual." *Philosophy and Science*, **33**:323–340, 1966.

[Whi90] Joe Whittaker. *Graphical Models in Applied Multivariate Statistics.* John Wiley & Sons, New York, 1990.

[Wri21] S. Wright. "Correlation and causation." *Journal of Agricultural Research*, **20**:557–585, 1921.